



A University of Sussex DPhil thesis

Available online via Sussex Research Online:

<http://sro.sussex.ac.uk/>

This thesis is protected by copyright which belongs to the author.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Please visit Sussex Research Online for more information and further details

***The computational analysis of
Post-translational modifications***

David Robert Damerell

**Thesis submitted for the degree of doctor of
philosophy**

September 2010

University of Sussex

Declaration:

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another university for award of any other degree.

Signed:

David R. Damerell

University of Sussex

David Robert Damerell

Thesis submitted for the degree of doctor of philosophy

The computational analysis of Post-Translational modifications

Summary:

The post translational modification (PTMs) of proteins presents a means to increase the proteome size and diversity of an organism through the inclusion of structural elements not encoded at the sequence-level alone. Their erroneous inclusion or exclusion has been linked to a variety of diseases and disorders thus their characterisation has the potential to present viable drug targets. The proliferation of newer high-throughput methods, such as mass spectrometry, to identify such modifications has led to a rapid increase in the number of databases and tools to display and analyse such vast amounts of data effectively. This study covers the development of one such tool; PTM Browser, and the construction of the underlying database that it is based upon. This new database was initially seeded with annotations from the Swiss-Prot and Phospho.ELM resources. The initial database of PTMs was then expanded to include a large repertoire of previously unannotated proteins for a selection of topical species (e.g. *Danio rerio* and *Tetraodon nigroviridis*). Orthologue assignments have also been added to the database – to allow for queries to be performed regarding the conservation of modifications between homologous proteins. The PTM Browser tool allows for a full exploration of this new database of PTMs – with a special focus on allowing users to identify modifications that are both shared between and are specific to particular species. This tool is freely available for non-commercial use at the following URL: <http://www.ptmbrowser.org>. An analysis is presented on the conservation of modifications between members of the tumour suppressor family, p53, using this new tool. This tool has also been used to analysis the conservation of modifications between super-kingdoms and Eukaryote species.

Acknowledgements

First and foremost I would like to thank my supervisor, Dr Sue Jones. She has been incredibly supportive since I undertook my masters in bioinformatics at Sussex University. I thank her for giving me the opportunity to work with her and the members of her group. Thanks to Beth Hellen, Ruth Spriggs, and Tim Beck for all the ideas and suggestions they provided whilst we were all at Sussex. My co-supervisor Dr Darren Thompson has also provided sound advice as I progressed through my project.

The work undertaken in this project would not have been possible without Jeremy Maris maintaining the network infrastructure and cluster resources upon which I became so dependent. I thank the Medical Research Council for funding me through the first three years of my DPhil project.

I am indebted to my parents, Tony and Jayne, who have emotionally and financially supported me for the last 8 years of my university education. Thanks to Sarah for always being just a phone call away, and for the support from Keron, Jessica and James. Andrew and Angela have been incredibly supportive to both me and my wife Claire.

Finally, my wife Claire has provided unending support for my research whilst undertaking her own at the same time. Without her support I would simply never have been able to undertake the project, and for that I am eternally grateful to her.

Contents

Chapter 1 Introduction 1

Section 1.1 The diversity of post-translation modifications:	1
Section 1.1.1 Phosphorylation	2
Section 1.1.2 Glycosylation	4
Section 1.1.3 Methylation	6
Section 1.1.4 Acetylation.....	6
Section 1.1.5 Lipidation	7
Section 1.1.6 Cleavage	10
Section 1.1.7 PTM Detection.....	10
Section 1.2 Bioinformatics resources.....	14
Section 1.2.1 Phosphorylation	16
Section 1.2.2 Glycosylation	16
Section 1.2.3 UniProtKB (http://uniprot.org).....	17
Section 1.2.4 dbPTM (http://dbptm.mbc.nctu.edu.tw).....	18
Section 1.3 Nomenclature.....	19
Section 1.3.1 RESID.....	19
Section 1.3.2 PSI-MOD	20
Section 1.4 Aims of thesis	21

Chapter 2 A new database of PTMs..... 24

Section 2.1 Summary.....	24
Section 2.2 Software and conventions.....	25
Section 2.3 Design approach	25
Section 2.4 Vocabulary.....	27
Section 2.4.1 Swiss-Prot vocabulary.....	28
Section 2.4.2 Primary PTMDB vocabulary.....	30
Section 2.4.3 PSI-MOD incorporation	35
Section 2.5 Annotation.....	39
Section 2.5.1 Swiss-Prot annotation format	39
Section 2.5.2 PTMDB annotation schema	40

Section 2.5.3 UniProtKB import into the PTMDB.....	46
Section 2.5.4 Phospho.ELM.....	52
Section 2.5.5 Negative PTM annotations.....	53
Section 2.6 Glycan structure import using the PDB2LINUCS tool.....	56
Section 2.6.1 Annotation extraction.....	58
Section 2.6.2 PTMDB import process.....	59
Section 2.6.3 Import into the PTMDB.....	60
Section 2.6.4 Glycan classification.....	61

Chapter 3 Incorporation of homology assignments into the PTMDB 66

Section 3.1 Summary.....	66
Section 3.2 Protein evolution.....	67
Section 3.2.1 Defining Homology.....	67
Section 3.2.2 Horizontal Gene Transfer (HGT).....	69
Section 3.2.3 Mixed ancestry.....	69
Section 3.2.4 Detection algorithms.....	69
Section 3.2.5 Summary.....	74
Section 3.3 CoPaO An implementation of the InParanoid algorithm.....	75
Section 3.3.1 Confidence value complications.....	77
Section 3.4 Orthologue detection in the PTMDB.....	77
Section 3.4.1 Species selection.....	78
Section 3.4.2 Redundancy in the UniProtKB.....	79
Section 3.4.3 Removing redundancy from the PTMDB.....	79
Section 3.5 Results.....	86
Section 3.5.1 Cluster Validation.....	86
Section 3.5.2 Proteome conservation.....	92
Section 3.6 Discussion.....	95

Chapter 4 Cross annotation of PTMs 97

Section 4.1 Summary.....	97
Section 4.2 Introduction.....	98
Section 4.2.1 Experimental determination.....	98
Section 4.2.2 Computational determination.....	99

Section 4.3 Cross-annotation protocol.....	104
Section 4.4 Results	107
Section 4.4.1 Incorporation of Pfam	107
Section 4.4.2 Target PTM set	108
Section 4.4.3 Cross-annotation results	114
Section 4.4.4 Comparison to UniProtKB release 2010_12	123
Section 4.5 Discussion	124
Chapter 5 PTM Browser	128
Section 5.1 Summary.....	128
Section 5.2 Application implementation details	129
Section 5.2.1 PTMDB inclusion.....	129
Section 5.2.2 Pfam complications	132
Section 5.3 Conservation analysis workflows	133
Section 5.3.1 Protein family analysis.....	133
Section 5.3.2 Taxonomic comparisons.....	139
Section 5.3.3 Generic conservation workflow	140
Section 5.4 Web service	154
Section 5.5 Discussion	157
Chapter 6 PTM Browser in action	158
Section 6.1 Summary.....	158
Section 6.2 Protein family analysis – p53.....	159
Section 6.2.1 p53 DNA binding domain (PF00870)	164
Section 6.2.2 p53 Transcriptional activation domain (PF08563)	166
Section 6.2.3 Discussion	169
Section 6.3 Taxonomic comparisons.....	172
Section 6.3.1 Conservation between super-kingdoms	172
Section 6.3.2 Conservation between Eukaryote species	175
Section 6.3.3 Discussion	179
Chapter 7 General Discussion.....	183
Section 7.1 PTM Databases	183
Section 7.2 Cross-annotation	185
Section 7.3 Proteomes and orthology detection	187

Section 7.4 PTM Conservation analysis tools.....	188
Section 7.5 Conservation analysis conclusions	190
Section 7.6 Biological relevance.....	193
Section 7.7 Future direction.....	194
References	196
Appendix.....	211
Appendix 1: p53 family UniProtKB list.....	211
Appendix 2: Source code	211

Figure 1: Common amino acids targeted for phosphorylation in both their native & modified states ...	2
Figure 2: Classification of N-Linked glycan structures.	5
Figure 3: Idealised 3D structures of Farnesyl and GeranylGeranyl.	8
Figure 4: Increase in PTM related publications found in PubMed between 1999 & 2009.	14
Figure 5: Intersection between PTM related resources and experimental determination techniques.	15
Figure 6: Entity relationship figure connections.	27
Figure 7: An idealised version of the Swiss-Prot PTM vocabulary schema	30
Figure 8: Swiss-Prot Glycosylation PTM type format.	32
Figure 9: Distribution of PTM types by PTM class in the PTMDB vocabulary.	33
Figure 10: PTMDB vocabulary schema diagram.	34
Figure 11: PTM type distribution in the PSI-MOD ontology.	37
Figure 12: PSI-MOD ontology schema diagram.	38
Figure 13: NCBI Entrez taxonomy schema used in the PTMDB.	42
Figure 14: PTMDB annotation storage schema.	44
Figure 15: PTMDB protein primary sequence validation routine.	45
Figure 16: Number of modified residues grouped by PTM class and evidence qualifier	50
Figure 17: Number of modified proteins grouped by PTM class and evidence qualifier	51
Figure 18: Distribution of negative PTM sites by PTM class.	56
Figure 19: Core N-linked glycan in CFG symbolic nomenclature and LINUCS format.	57
Figure 20: N-linked glycan core types used in the classification of structures from the PDB.	61
Figure 21: Glycan classification support in the PTMDB vocabulary schema	63
Figure 22: Simple model of a gene undergoing duplication and subsequent speciation	68
Figure 23: InParanoid confidence value equation for in-paralogues.	73
Figure 24: Alternative confidence value equation for second level in- paralogues.	77
Figure 25: Redundancy removal protocol.	81
Figure 26: Comparison of model-based prediction with cross-annotation.	101
Figure 27: Prosite, tyrosine kinase phosphorylation site pattern (PDOC00007).	102
Figure 28: Limited conservation of N and C termini of proteins.	109
Figure 29: Distribution & annotation of Prenylation acceptor sites in the Pfam database.	110
Figure 30: Percentage of GPI-anchored & phosphorylation sites found in Pfam domains.	111
Figure 31: Conservation of the regions surrounding homologous PTM sites.	115
Figure 32: Sequence window analysis of the Swiss-Prot cross-annotation set.	116
Figure 33: Distribution of predicted Glycosylated, N-linked Glycosylated and Phosphorylated proteins between different Eukaryote species.	120
Figure 34: Distribution of phosphorylation cross-annotations between the different types of phosphorylation, grouped by species.	121
Figure 35: PTM Browser client web interface	130
Figure 36: Domain layout of the protein IMDH1_HUMAN (SwissProt ID).	133
Figure 37: PTM Browser protein family workflow.	137
Figure 38: PTM Browser accession entry component.	138
Figure 39: PTM Browser conservation table format	138
Figure 40: PTM Browser Pfam domain starred alignment example.	139
Figure 41: PTM Browser conservation results panel	139
Figure 42: PTM Browser taxonomic comparison workflow.	142
Figure 43: PTM Browser add experiment user interface.	143
Figure 44: PTM Browser experiment component.	144
Figure 45: PTM Browser query selection panel.	145
Figure 46: PTM Browser complement experiment example.	148
Figure 47: Intersect summary performed between two species.	148
Figure 48: PTM Browser example intersection graph	149

Figure 49: PTM Browser experiment browser interface.....	151
Figure 50: PTM Browser field restriction dialog box.	151
Figure 51: PTM Browser CoPaO cluster view.	153
Figure 52: PTM Browser API view.....	156
Figure 53: Domain structure of an idealised p53 (a) and p63 (b) protein.....	160
Figure 54: p53 family phylogenetic tree.	162
Figure 55: Alignment of Ser-99(4) region from various members of the p53 family.	165
Figure 56: Alignment of Ser-215(171) region from various members of the p53 family.	166
Figure 57: Conservation of modifications in the p53 transcriptional activation domain.	168
Figure 58: Pfam domains with experimentally verified PTM annotations conserved between Bacterial and Eukaryotic proteins.....	172
Figure 59: Number of modified residues conserved between Eukaryotic and Bacterial species	173
Figure 60: Alignment of a conserved N-linked glycosylation site in the Pfam domain PF00082	174
Figure 61: Pfam domains with conserved PTMs between Eukaryotic and Archeal species s.	174
Figure 62: Modification sites conserved across all three super-kingdomss.	175

Table 1: List of PTM classes..	1
Table 2: PSI-MOD relationships	21
Table 3: PSI-MOD relationship examples	21
Table 4: UniProtKB feature keys associated with PTM types	29
Table 5: UniProtKB PTM vocabulary attributes supported by the PTMDB	30
Table 6: PSI-MOD term attributes supported by the PTMDB..	35
Table 7: Swiss-Prot PTM feature table example entries	39
Table 8: Non-PTM annotation related UniProtKB fields imported into the PTMDB.	46
Table 9: Distributions of PTM annotations imported from Swiss-Prot grouped by PTM class..	48
Table 10: Distribution of Swiss-Prot PTM annotated sites grouped by PTM class and evidence.	49
Table 11: Distribution of Swiss-Prot PTM annotated proteins grouped by PTM class and evidence	49
Table 12: Breakdown of the reasons for Phospho.ELM annotation rejection.	53
Table 13: Phospho.ELM import statistics.	54
Table 14: Additional PTM types created for the Swiss-Prot negative PTM annotations	55
Table 15: Comparison of tools that can extract glycan structures from the PDB.	57
Table 16: Glycan classification in the PDB2LINUCS dataset imported into the PTMDB.	64
Table 17: Intersect between the Swiss-Prot and PDB2LINUCS imported Glycosylation annotations.	65
Table 18: Summary statistics for PDB2LINUCS glycan annotated TrEMBL entries	65
Table 19: Orthologue detection techniques	70
Table 20: Species selected for orthologue detection using the CoPaO program.	78
Table 21: Distribution of species included in the proteome generation procedure, by Super-kingdom	83
Table 22: Number of proteins removed at each step of the redundancy removal protocol	84
Table 23: Estimated proteome coverage in the PTMDB proteome sets	85
Table 24: Proteome coverage in the PTMDB for selected model species & species of special interest.	86
Table 25: Comparison of Homo sapiens and Mus musculus orthologue sets in the PTMDB and InParanoid 6.1 datasets	87
Table 26: RSS validation of orthologue clusters	91
Table 27: Percentage of the Homo sapiens proteome which is orthologous to other species.	94
Table 28: Comparison between the likelihoods of a residue, at each position in a protein, being either (A) GPI-anchored or (B) phosphorylated, being annotated in a Pfam domain.	112
Table 29: The percentage of proteins that have at least one of their known acceptor sites, for the given PTM class, in a Pfam domain.	113
Table 30: Percentage of acceptor sites in a Pfam domain.	114
Table 31: Gross number of cross-annotations made in the Swiss-Prot & TrEMBL database	117
Table 32: Number of proteins with cross-annotations for each PTM class.	118
Table 33: Bacteria cross-annotation statistics	119
Table 34: Eukaryote cross-annotation statistics.	121
Table 35: Inter-super-kingdom predictions	123
Table 36: Comparison of cross-annotations in the PTMDB, and PTM annotations in UniProtKB release 2010_12	124
Table 37: PTM Browser PTMDB support	131
Table 38: PTM Browser PTMDB modified table structure.	132
Table 39: Example annotation set for an experiment.	146
Table 40: List of the unique combinations of the fields Pfam Accession and Aln Position observed in the annotation set shown in Table 39	146
Table 41: The conservation of PTMs in the p53 TA and DNA binding domains.	163
Table 42: Conservation of modified proteins between H. sapiens and of other Eukaryotes species	177
Table 43: Conservation of modified domain residues between H. sapiens & M. musculus	179

Abbreviations

BBH	<u>B</u> last <u>B</u> est <u>H</u> it
COG	<u>C</u> lusters of <u>O</u> rthologous <u>G</u> roups of proteins
CoPaO	<u>C</u> lusters of <u>P</u> aralogues and <u>O</u> rthologues
CV	<u>C</u> ontrolled <u>V</u> ocabulary
DAG	<u>D</u> irected <u>A</u> cylic <u>G</u> raph
GO	<u>G</u> ene <u>O</u> ntology
GPI	<u>G</u> lycosylphosphatidylinositolylation
HGT	<u>H</u> orizontal <u>G</u> ene <u>T</u> ransfer
HPC	<u>H</u> igh <u>P</u> erformance <u>C</u> omputing
HUGO	<u>H</u> uman <u>G</u> enome <u>O</u> rganisation
HUPO	<u>H</u> uman <u>P</u> roteome <u>O</u> rganisation
KEGG	<u>K</u> yoto <u>E</u> ncyclopaedia of <u>G</u> enes and <u>G</u> enomes
KOG	<u>E</u> karyotic <u>O</u> rthologous <u>G</u> roups
LCA	<u>L</u> ast <u>C</u> ommon <u>A</u> ncessor
MRCA	<u>M</u> ost <u>R</u> ecent <u>C</u> ommon <u>A</u> ncessor
MS	<u>M</u> ass <u>S</u> pectrometry
OMIM	<u>O</u> nlne <u>M</u> endelian <u>I</u> nheritance in <u>M</u> an
PDB	<u>P</u> rotein <u>D</u> atabank
PSI	<u>P</u> ercentage <u>S</u> equence <u>I</u> dentify
PTM	<u>P</u> ost <u>T</u> ranslational <u>M</u> odification
PTMDB	<u>P</u> TM <u>D</u> atabase
Ras	<u>R</u> at <u>S</u> arcoma
RDBMS	<u>R</u> elational <u>D</u> atabase <u>M</u> anagement <u>S</u> ystem
RSS	<u>R</u> elative <u>S</u> pecificity <u>S</u> imilarity
SI	<u>S</u> equence <u>I</u> dentify
TrEMBL	<u>T</u> ranslated <u>E</u> MBL
UniProt	<u>U</u> niversal <u>P</u> rotein Resource
UniProtKB	<u>U</u> niprot <u>K</u> nowledge <u>B</u> ase
PHOSIDA	(<u>P</u> hosphorylation <u>S</u> ite <u>D</u> atabase)
PSSM	<u>P</u> osition <u>S</u> pecific <u>S</u> coring <u>M</u> atrix
MDD	<u>M</u> aximum <u>D</u> ependence <u>D</u> ecomposition
NET	<u>N</u> CBi <u>E</u> ntrez <u>T</u> axonomy

Chapter 1

Introduction

This chapter begins with an overview of the diversity of PTMs (Post-translational modifications) with particular emphasis on the important functions that they play in protein: structure; function and regulation. This is followed by a review of the experimental detection techniques that are commonly used to identify PTMs. The second half of this chapter discusses current PTM databases and the nomenclature that is used in them to represent PTMs.

Section 1.1 *The diversity of post-translation modifications:*

It is a common statement that humans have 98% sequence similarity with the chimp genome, leaving a miniscule 2% to account for the differences between us. This figure is however based purely upon sequence-level comparison, and therefore does not take account of differences in gene expression, alternate splicing and post-translational modification. The PTM of a protein is any form of structural alteration that occurs at the post-translational level. This provides an extra degree of proteome diversity and complexity, which is not determined by the primary sequence alone. There are a great number of different PTM classes – 24 of which are shown in Table 1. A small selection of these PTMs will now be reviewed in detail.

N-linked Glycosylation	O-linked Glycosylation	C-linked Glycosylation
S-linked Glycosylation	Phosphorylation	Methylation
Hydroxylation	Acetylation	Palmitoylation
Myristoylation	Prenylation	Formylation
Amidation	GPI-anchor	Sulfation
ADP-Ribosylation	Citrullination	Bromination
S-nitrosylation	Glutathionylation	Oxidation
Iodination	Disulphide bonds	Cleavage

Table 1: List of main PTM types. This list is in-part based on those modification classes stored in the Swiss-Prot database (see Table 9 for the complete list).

Section 1.1.1 Phosphorylation

Phosphate is one of the smallest ligands that can be used to modify a protein but is perhaps the most abundant form of PTM. Phosphorylation is carried out by protein kinases, which are highly residue-specific. The protein kinases are commonly associated with cell signalling cascades, however the serine/threonine-specific protein kinase is not the only type that exists in nature, in eukaryotes the other main type is tyrosine-specific and in bacteria it is histidine that is most commonly phosphorylated for this purpose (Choi *et al.* 2008; Khorchid and Ikura 2006). The addition of phosphate to the hydroxyl group of serine, threonine and tyrosine produces a stable phosphoester (P-O) bond, whereas the phosphorylation of histidine results in an acid-labile phosphoramidate (P-N) bond (Attwood *et al.* 2007). Figure 1 shows these amino acids in their native and modified forms.

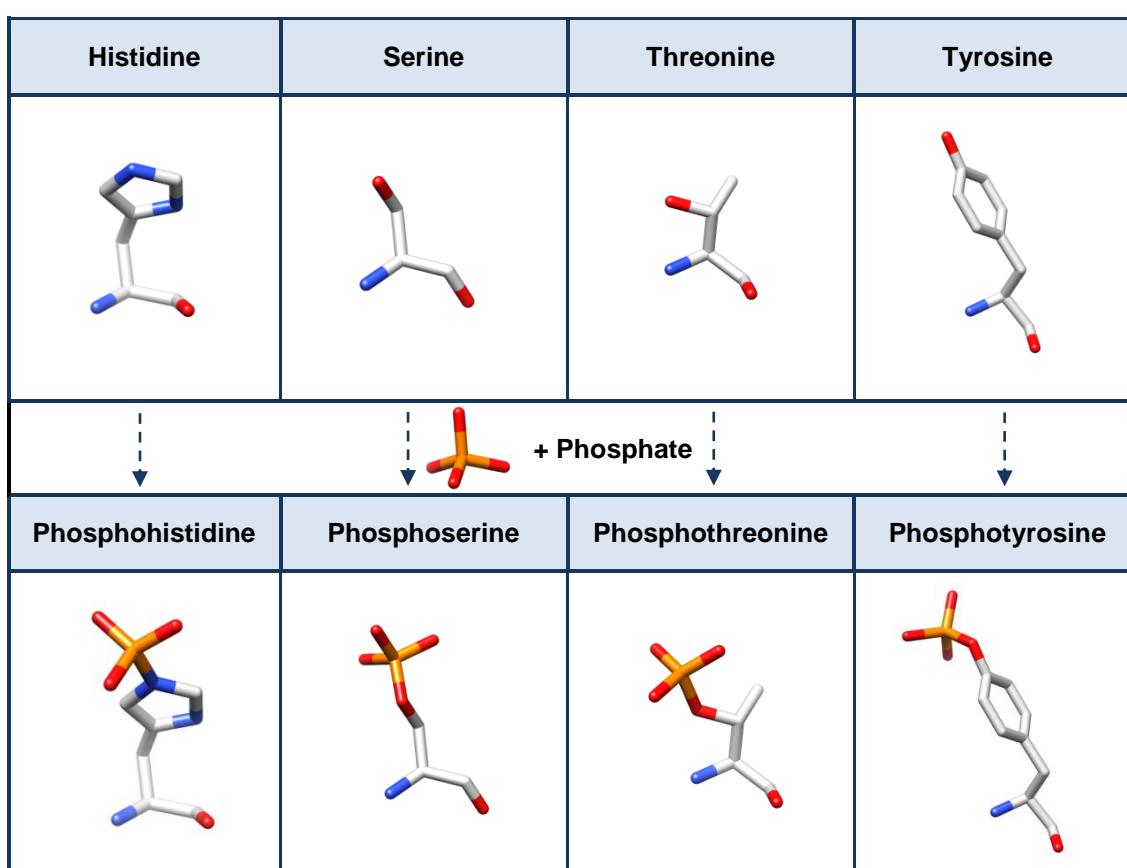


Figure 1: Common amino acids targeted for phosphorylation in both their native and modified states. The amino acids shown are from the following PDB entries: phosphoserine 3NAY (Nagashima *et al.* 2011), phosphotyrosine 3N8M (Delorbe *et al.* 2010), phosphothreonine 3MY1 (Baumli, Endicott, and Johnson 2010) and N3-phosphohistidine 1QWO (Xiang *et al.* 2004).

The reason for the difference in phosphorelay preferences between these groups is not fully understood, however Choi *et al.* postulate that the serine/threonine/tyrosine multi-step system may have presented an evolutionary advantage due to its ability to incorporate different signal sources as well as the ability to tightly control the duration of the signalling burst (Choi *et al.* 2008). Whereas Attwood *et al.* suggest that the propensity for hydrolysis of the phosphoramidate bond would present an evolutionary advantage for systems requiring a rapid 'on/off switch', by removing the requirement for additional components to terminate the signal (Attwood *et al.* 2007). These hypotheses would support the difference in amino acid preference between the distinct superkingdoms, with eukaryotes requiring complex patterning of gene expression and precise timing, and prokaryotes requiring a rapid response to changes in their extracellular environment. Phosphorylation however is not limited to these four amino acids, arginine (Besant, Attwood, and Piggott 2009), lysine (Besant, Attwood, and Piggott 2009), cysteine (McAdams *et al.* 2008), aspartate (Attwood, Besant, and Piggott 2010), glutamate (Attwood, Besant, and Piggott 2010) and the already post-translationally modified residue hydroxyproline (Kühlberg, Haid, and Metzger 2010) are also potential phosphorylation targets.

Hyperphosphorylation of p53 has been observed in many tumour derived cell lines (Minamoto *et al.* 2001). The tumour suppressor protein Rb (Retinoblastoma protein) has been shown to play an important role in controlling the G1-S checkpoint (Harb *et al.* 2009). In its unphosphorylated form Rb is bound to the transcription factor E2F (Chellappan *et al.* 1991). The phosphorylation of Rb results in the release of the E2F transcription factor and subsequent cell cycle progression (Adams 2001). Research has been published that suggests that Silibinin (available as a dietary supplement) promotes the formation of the unphosphorylated form of Rb, which may have therapeutic and preventive properties in the treatment of prostate cancer (Tyagi, Agarwal, and Agarwal 2002).

Section 1.1.2 Glycosylation

Glycosylation is generally split between two main types: N-linked and O-linked, defined by the functional group of the amino acid to which the sugar is attached (Spiro 2002). N-linked glycosylation is associated with either asparagine or arginine, which contain an amide or amino functional group, whereas O-linked glycosylation is associated with hydroxyl-containing amino acids, e.g. serine, threonine and tyrosine (Spiro 2002). The N-glycosidic bonds are best characterised by the β -glycosylamine linkage between asparagines and N-acetylglucosamine (GlcNAc) (Spiro 2002).

N-linked glycans are characterised by a core penta-saccharide formed of two GlcNAc residues and three mannose (Man) residues. A series of glycosyltransferases and hydrolyases are responsible for both adding and removing residues from the core-penta-saccharide (Walsh 2006). N-linked glycans are usually classified into the following groups: high mannose, complex, core and truncated (Walsh 2006) (see Table 2 for example structures and Section 2.6.4 for further information on N-linked glycan classification).

Glycosylation was until fairly recently considered to be a PTM that was specific to Eukaryotes. This view was changed with the discovery that *Campylobacter jejuni* (a bacterial species) is able to N-link glycosylate a number of its proteins (including flagella) (Szymanski *et al.* 1999). Archaea have also been shown to glycosylate their proteins – in fact in a recent review on the subject it was highlighted that archaeal glycosylation is more prevalent than bacterial (Yurist-Doutsch *et al.* 2008).

In recent years the link between differences in glycosylation patterns and disease has become an area of great interest (Hakomori 2002; Jaeken and Matthijs 2007). Interestingly all forms of human cancers have glycosylation defects, though it is not yet clear whether this is a cause or a symptom (Hakomori 2002). Alteration of these modifications is potentially linked to metastasis by changing the way in which a cell can interact with those around it, as well as inhibiting signalling cascades that would ultimately lead to apoptosis (Hakomori 2002).

Glycoproteins are an extremely diverse group, which are closely associated with cell signalling. Glycosylation is involved in the presentation of self antigens to the immune system, via MHC (Parham 1996) and is used to distinguish non-self (Dziarski 2003). Perhaps one of the most well-known effects of differences in the glycosylation of cells is the ABO blood grouping system in humans (Takasaki, Yamashita, and Kobata 1978). Erythrocytes from all four blood groups: A, B, O and AB, contain the H chain displayed on their cell surface (Yamamoto *et al.* 1990). This can be further extended by glycosyltransferases, the precise function of which varies to produce the different groups, e.g. blood group A is produced by α 1-3 N-acetylgalactosaminyltransferase (A-transferase), whereas group B is produced by α 1-3 galactosyltransferase (B-transferase) (Yamamoto *et al.* 1990). Blood group O is produced owing to the lack of a functional transferase thus contains the H-chain alone (Yamamoto *et al.* 1990). Both functional glycosyltransferases are the product of the ABO gene on chromosome 9. However, the substrate specificity is altered by single nucleotide polymorphisms between individuals (Yazer and Olsson 2008).

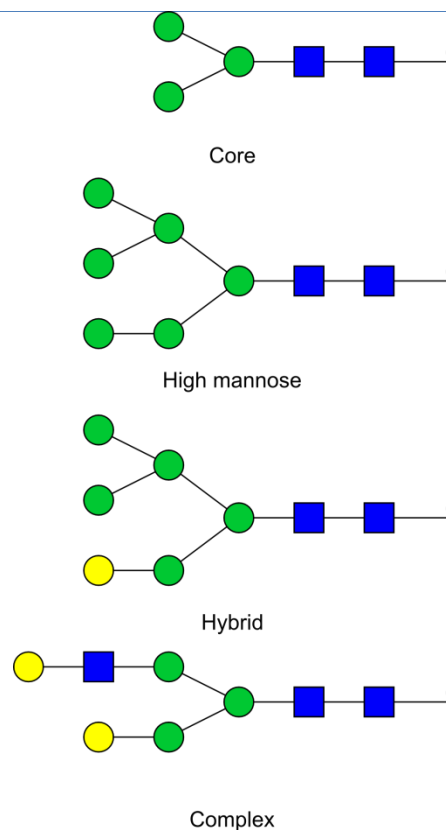


Figure 2: N-Linked glycan classification examples. Images created using GlycoWorkbench (Ceroni *et al.* 2008).

Section 1.1.3 Methylation

Protein methylation involves the transfer of a methyl (CH_3) group from the substrate S-adenosylmethionine (SAM) onto a number of different amino acid side chains (Clarke 1993). The two most common forms of methylation are the methylation of carboxyl groups and nitrogen atoms (Clarke 1993). N-methylation is an irreversible reaction (Clarke 1993) that commonly occurs at arginine and lysine residues (Grillo and Colombatto 2005). O-methylation is a reversible reaction (Clarke 1993) that has been observed on leucine (Bryant, Westphal, and Wadzinski 1999) and asparagine residues (Klotz and Glazer 1987).

One of the first examples of O-methylation was observed on the photosynthetic accessory Phycobiliproteins (Klotz and Glazer 1987). S-methylation, the methylation of cysteine residues, was first reported from an MS/MS (tandem Mass Spectrometry) analysis of the proteins found in the eye lens (Lapko, Smith, and Smith 2002). C-methylation has also been observed at the C-2 position of glutamine and C-5 position of arginine in the archeal Methyl-coenzyme M Reductase (Selmer *et al.* 2000).

Ha-ras (a key oncogene) is both prenylated (Section 1.1.5(b) and methylated, both of which are thought to be required for membrane localisation (Clarke *et al.* 1988). It appears that one of the effects of the chemotherapeutic agent Methotrexate is to reduce O-methylation of Ras (Rat Sarcoma) family members (Philips 2004).

Section 1.1.4 Acetylation

Proteins can either be acetylated at the ϵ -amino group of lysine residues or α -amino group of N-terminal methionine (or preceding residue if the methionine has been removed) (Soppa 2010). Acetylation has been observed in all three superkingdoms (Soppa 2010). N-terminal methionine acetylation is a co-translationally process carried out by N-terminal acetyltransferases (NATs) (Polevoda *et al.* 1999). N-terminal acetylation may act as a degradation signal in the ubiquitin dependent degradation pathway (Hwang, Shemorry, and Varshavsky 2010).

Acetylation of histone proteins has been known since at least 1976 (Taylor and Cook 1981). Acetylation of lysine residues on Histone tails (by histone acetyltransferases, HATs) is associated with the activation of genes and deacetylation (by histone deacetylases, HDACs) with the repression of genes (via chromatin remodelling) (Eberhartner and Becker 2002). Histone acetylation is associated with a permissive (supporting transcription) chromatin structure (Eberhartner and Becker 2002). Histones H3 and H4 located at the 5' end of FMR1 (fragile X mental retardation 1) are acetylated in cells derived from healthy individuals; in contrast cells from fragile X patients show decreased acetylation (Coffee *et al.* 1999).

Section 1.1.5 Lipidation

Section 1.1.5(a) *Glycosylphosphatidylinositolylation*

The glycosylphosphatidylinositol (GPI)-anchors are a way to anchor a protein to the cell membrane (Ferguson and Williams 1988). As the name would indicate they are formed of both glycosyl and lipid components, with the hydrophobic lipid chains facilitating membrane anchoring and the glycans providing docking sites for protein molecules (Ferguson and Williams 1988). The synthesis of GPI-anchored proteins is complex, involving a flippase - required to transport the phosphatidylinositol base structure into the lumen of the ER for further extension, as well as the Golgi apparatus, which facilitates membrane insertion (Almeida, Layton, and Karadimitris 2009). As with other PTMs, its disruption is linked to disease, namely inherited GPI-deficiency and paroxysmal nocturnal hemoglobinuria (acquired) that results in thrombosis and haemolytic anaemia, respectively (Almeida, Layton, and Karadimitris 2009).

Section 1.1.5(b) *Prenylation*

Prenylation involves the transfer of either the 15-carbon Farnesyl ($C_{15}H_{25}$) group or 20-carbon ($C_{20}H_{33}$) GeranylGeranyl group (see Figure 3) to c-terminal cysteine residues (Roskoski 2003). Farnesyl groups are transferred by Farnesyl transferases (FTases) and GeranylGeranyl groups by GeranylGeranyl transferases (GGTases) I and II (Roskoski 2003). FTase and GGTase I cleave after the cysteine residue in the motif CaaX (where "a" is any aliphatic residue and X is any amino acid) and form a thioether bond to the Farnesyl or

GeranylGeranyl groups (Maurer-Stroh and Eisenhaber 2005). GGTaseII recognises the motifs CC and CxC (Maurer-Stroh and Eisenhaber 2005).

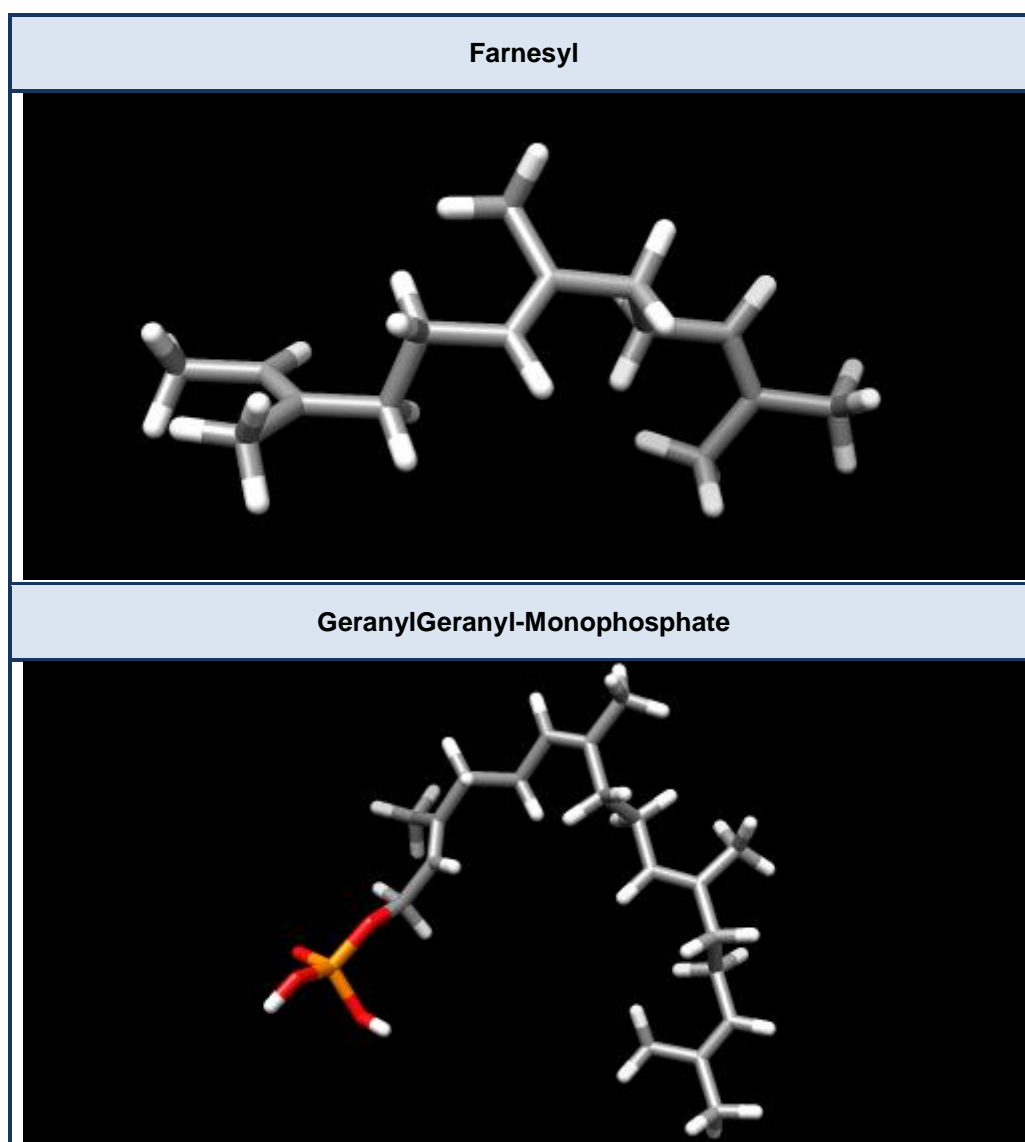


Figure 3: Farnesyl and GeranylGeranyl idealised 3D structures. PDB coordinates were obtained from the PDB Ligand Expo database (Feng et al. 2004).

Like other lipid modifications prenylation has been shown to be required for the membrane association of a number of proteins. For example the double prenylation (GeranylGeranyl) of the GTPase Rab family of proteins (belonging to the Ras superfamily) has been shown to be essential for their correct localisation (Calero *et al.* 2003). Ras family members are key regulators of cell proliferation and are frequently mutated in *H. sapiens* tumours (Downward 2003). As the prenylation of Ras family members is essential for their function, FTases and GGTases were obvious drug targets (Downward 2003). In fact research has shown that Farnesyltransferase and GeranylGeranyltransferase

inhibitors (FTIs and GGTIs) may be relevant as therapeutic agents in the treatment of both cancer and viral related diseases. For example the FTIs BZA-5B and FTI-277 have been shown to block the proliferation of Hepatitis delta virus (HDV) (a cause of chronic and acute liver disease) (Bordier *et al.* 2002).

Section 1.1.5(c) Palmitoylation and Myristoylation

Palmitoylation involves the formation of either an amide (n-linked) or thioester bond (s-linked) between a long chain fatty acid (mainly palmitate) and a cysteine residue (Linder and Deschenes 2003). N-myristoylation involves the formation of a covalent bond between the n-terminal of a glycine residue and the 14 carbon saturated fatty acid, myristate by the enzyme N-myristoyltransferase (NMT) (Wright *et al.* 2009). S-palmitoylation is a reversible PTM whereas N-myristoylation is not (Linder and Deschenes 2003).

S-palmitoylation is required to localise the protease BACE1 (of which the Amyloid precursor protein (APP) is a target) to lipid rafts (Vetrivel *et al.* 2009). There is conflicting evidence regarding the importance of S-palmitoylation in the eventual deposition of β -amyloid peptides ($A\beta$) in the brain (formed by the cleavage of the APP protein) (Meckler *et al.* 2010). Whilst the cleavage of APP by BACE1 and γ -secretases into $A\beta$ has been shown to be independent of their S-palmitoylation status *in vitro*, research suggests this may not be the case *in vivo* (Vetrivel *et al.* 2009; Meckler *et al.* 2010).

Many viruses are dependent on the host NMT, including HIV-1, for replication, which has led to the development of N-myristoylation inhibitors. For example, N-myristoylation of the Z protein, belonging to Lassa virus (LASV), has been shown to be essential for its recruitment to the plasma membrane (Strecker *et al.* 2006). Treatment of LASV infected cells with N-myristoylation inhibitors has been shown to reduce the replication efficiency of the virus (Strecker *et al.* 2006). N-myristoylation has also been implicated in affecting the stability of proteins. For example the catalytic subunit of cAMP-dependent protein kinase is stabilised after being N-myristoylated (Yonemoto, McGlone, and Taylor 1993). The non-receptor tyrosine kinase family, Src members are key regulators of cellular signalling, for which N-myristoylation is essential for their

regulation and membrane association (Patwardhan and Resh 2010). NMT has also been found to be overexpressed in *H. sapiens* colorectal adenocarcinomas (Raju, Moyana, and Sharma 1997).

Section 1.1.6 Cleavage

The first protein to have its complete amino acid sequence determined was insulin, a discovery that won Frederick Sanger the 1958 Nobel Prize in Chemistry (Sanger and Thompson 1953a; Sanger and Thompson 1953b; Sanger and Tuppy 1951a; Sanger and Tuppy 1951b). Insulin however is subject to two forms of post-translational modification: proteolytic cleavage and disulphide bond formation, which are required to convert proinsulin, as expressed from the INS gene, to the active hormone insulin (Ryle *et al.* 1955; Steiner and Oyer 1967). Proteolytic cleavage is a frequent occurrence amongst secreted proteins, owing to the requirement for removal of the export signal peptide, thus it is commonly associated with hormones, e.g. hGH, somatostatin, calcitonin, glucagon and parathyroid (Rholam, Nicolas, and Cohen 1986). Likewise disulphide-bridges are also regularly linked to exported proteins in both eukaryotes and bacteria (Sevier and Kaiser 2002).

Section 1.1.7 PTM Detection

It is standard practice for anyone expressing or purifying a protein to analyse their samples by PAGE (polyacrylamide gel electrophoresis) to get an indication of purity, yield and size of protein (Steinberg 2009). A polyacrylamide gel is formed of a matrix of cross-linked acrylamide monomers. The cross-linked nature of the acrylamide forms pores that serve to separate proteins based on size, with pore size dependent on acrylamide concentration. SDS (sodium dodecyl sulphate) is an anionic detergent that is commonly added to PAGE gels (SDS-PAGE), which is able to denature proteins and mask their natural charge (Wilson 2000). SDS-polyacrylamide gels are referred to as denaturing gels and polyacrylamide gels as native gels (because the protein runs in its native conformation in the absence of SDS) (Wilson 2000). Thus the benefit of each is as their names suggest, native gels allow you to study the natural state of the protein and denaturing gels allow you to study the peptide chain (Wilson 2000). In one dimensional (1D) polyacrylamide gel electrophoresis proteins are

separated based on their ability to migrate through the gel, generally resulting in a size gradient with larger proteins running more slowly although native gels vary in this point due to the different folds and charges like-sized proteins can have (Wilson 2000).

In order to separate by an additional dimension (2D) the 1D gel, or specific lane, is rotated 90°C and run again, usually with a pH gradient applied in order to separate by the isoelectric point (pI) (Wilson 2000). The benefit of this method is that two proteins of equal size are unlikely to share a pI thus a 2D gel could distinguish between them when a 1D gel could not.

These methods are useful for separating unmodified proteins but in order to determine whether a PTM has been added additional steps are required. Western blots can be performed that make use of PTM specific antibodies that act as reporters for particular modifications (Miller, Crawford, and Gianazza 2006). In recent years the use of antibodies to detect PTMs has grown in popularity with many pharmaceutical companies (e.g. Millipore, Perkin Elmer and Sigma-Aldrich) producing a range of products and kits to meet the high demand. As with unmodified proteins many companies also provide an antibody generation service when a generic antibody is not appropriate. PTM-specific antibodies can be produced that are specific to particular proteins (i.e. they recognise part of the modification and other parts of the proteins 3D structure). This allows for protein specific-PTM antibodies to be used with mixed population samples. As with all antibodies there can however be cross-reactivity (Fuchs *et al.* 2011). For example a recent study by Fuchs *et al.* 2011 demonstrated that antibodies raised against di- and tri-methyl-lysine modifications at positions 4 and 79 of histone protein H3, were cross-reactive with other di- and tri-methyl-lysine modifications. The same study also demonstrated that adjacent modifications can affect the ability of antibodies to recognise the modifications they were raised against.

A common way of detecting proteins that are modified is to identify spots on protein gels that change position in the presence and absence of an inhibitor of a particular modification. For example N-terminal α -amine acetylation has traditionally been detected by comparing the position of protein spots on 2D

SDS gels in the presence and absence of N-terminal acetyltransferases (Van Damme *et al.* 2009). Although recently gel-free techniques have been developed that both separate and enrich N-terminal α -amine acetylated peptides for detection using MS (Van Damme *et al.* 2009).

Modified proteins can also be identified on gels through the radiolabelling of PTMs. For instance phosphorylated proteins can be detected by growing cells on a medium that contains radiolabelled γ [^{32}P]-ATP (this is only possible because no amino acid includes phosphate) (Miller, Crawford, and Gianazza 2006). Phosphoserine and phosphothreonine can also be detected on protein gels using Methyl green after the phosphates have been hydrolysed with NaOH (Miller, Crawford, and Gianazza 2006). Fluorescent stains have also been created that can stain proteins with particular modifications (Miller, Crawford, and Gianazza 2006).

MS can be used to identify both the position and type of virtually all PTMs. Traditional proteomic experiments start with the digestion of proteins with an enzyme such as trypsin. The resulting mixture of peptides is then usually separated using a HPLC (High Performance Liquid Chromatography) system that is connected to a mass spectrometer (commonly referred to as on-line LC-MS). The mass spectrometer will record a peak for each observed peptide fragment with the corresponding mass to charge ratio (m/z). Software is then used that analyses the resulting mass spectrum to identify which proteins were present in the sample. Most algorithms start with a database of protein sequences, which are digested *in-silico* with the same enzyme used in the real experiment. The software then deduces the m/z of each peptide fragment based on the conditions of the original experiment (i.e. negative or positive ion mode). This provides the software with a complete list of m/z values for each protein. Algorithms tend to indicate which proteins are present in a sample based on the number of expected fragments for a particular protein that were actually observed in the MS spectra. PTMs can be detected by looking for peaks that have an m/z shift that is equivalent to an *in-silico* generated peptide fragment having a particular modification. Even if a modified fragment is identified it does not necessarily follow that the software will be able to unambiguously tell the user which residue was modified in the fragment. For

example, a phospho-peptide may contain multiple residues that could carry the phosphate group.

The MS analysis of glycoproteins is a particularly challenging area of research due to the complex nature of glycan structures. One simple method of identifying glycopeptides in a mixed population is to use lectin affinity chromatography (e.g. by using Concanavalin A, which recognises the trimannosyl-core of N-linked glycans) (Fan *et al.* 2004; Brewer and Bhattacharyya 1986). N-linked glycans are commonly removed from the Asn residues of glycopeptides using PNGaseF. This enzyme has the effect of converting the modified Asn residues to Asp (Fan *et al.* 2004). PNGaseF treatment can therefore be used as an indirect method of identifying glycopeptides in a mixed sample (by the corresponding m/z shift). Identifying both the site and corresponding glycan structure is a particularly difficult area of research (An, Froehlich, and Lebrilla 2009). One of the problems is that glycans fragment at much lower collision energies than peptides (An, Froehlich, and Lebrilla 2009). Technologies are however under development that allows for the acquisition of glycan and peptide fragments. For example 157nm light has been shown to produce glycan and peptide fragments (including glycan cross-ring cleavages, which are required to obtain linkage information) (Zhang and Reilly 2009). Deducing which glycan structures produced a particular set of peaks is far from a straight forward process. This, in part, stems from the large number of ways in which monosaccharides can be joined together. Glycans typically need to undergo multiple rounds of fragmentation (e.g. MS/MS, MS/MS/MS, ... MS^n) before software is able to deduce which structures are present (i.e. fragmenting fragments) (Ceroni *et al.* 2008). Glycoinformatic tools have been designed that attempt to make glycan structure determination a semi-automated process. For instance GlycoWorkbench is able to fragment user supplied structures and automatically annotate a given spectra (Ceroni *et al.* 2008). Alternatively this tool can function in a similar way to general proteomics software, whereby it carries out an *in silico* fragmentation of a database of known glycan structures (Ceroni *et al.* 2008). Other tools have been designed that reduce the glycan structure search space to only those structures that can be produced by known biosynthetic pathways (Goldberg *et al.* 2005).

Section 1.2 *Bioinformatics resources*

With such a diverse set of post-translational modifications identified to date it should come as no surprise that a large number of databases, and associated tools, have been created. Modern proteomics techniques are promising to deliver an unprecedented volume of data, including the identification of countless new sites of PTM (Prince *et al.* 2004). Figure 4 demonstrates the explosion that has occurred in published papers associated with keywords connected to the analysis and detection of PTMs.

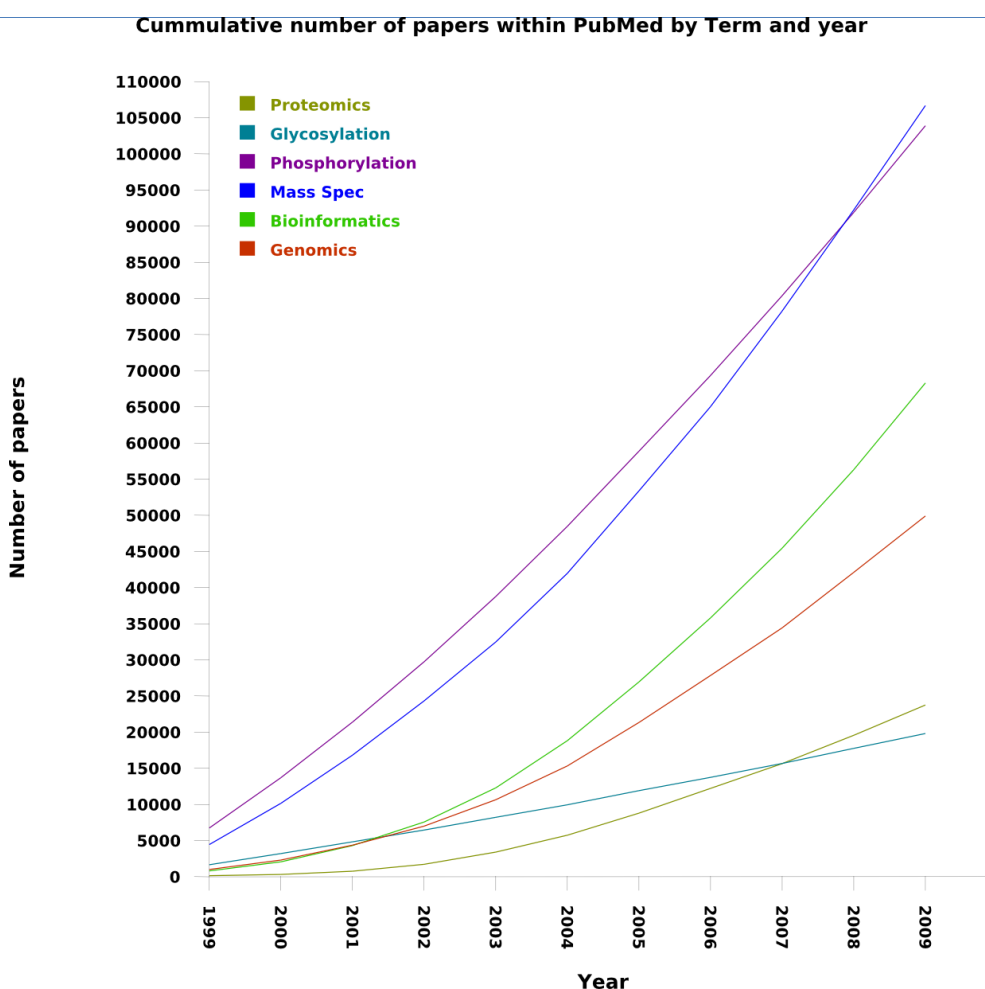


Figure 4: Cumulative number of papers in PubMed that are associated with specific PTM related terms between 1999 and 2009. This graph was generated using the class `org.drd20.bioinformatics.database.PubMed`, which is part of the Bioinformatics Basic Perl (BBP) project (created by the author). More information regarding the BBP project can be found in Appendix 2. The PubMed class utilizes the web service API provided for <http://ncbi.nlm.nih.gov> resources called EUtils (<http://eutils.ncbi.nlm.nih.gov>). The ESearch command of this API is used by this class to obtain counts for the number of papers published between specific dates that are associated with a specific term. Usage of the PubMed class can be found here: http://wiki.ptmbrowser.org/index.php/BBP_PubMed_Graph. The PubMed class utilizes the class `org.drd20.core.graphing.svg.SVGBarChart2` to produce graphs of the retrieved results. SVGBarChart2 renders graphs using the Cairo vector graphics library.

Bioinformaticians have successfully dealt with similar influxes of data from the fields of genomics and transcriptomics during the past 10 years (Prince *et al.* 2004). Challenges facing the new field of Proteomic Informatics include: data storage and availability, publication standards, and data analysis techniques (Prince *et al.* 2004). Figure 5 displays the bioinformatics resources that have been created to aid in the analysis and storage of PTM annotations – set in the context of the experimental techniques that are used to detect them. The experimental and computational techniques that can be used to identify PTMs are reviewed in Chapter 4. Chapter 2 documents the creation of a new database of PTMs in detail. What is now presented is a review of the existing databases and interfaces that they provide, which should be used to place this new database into context.

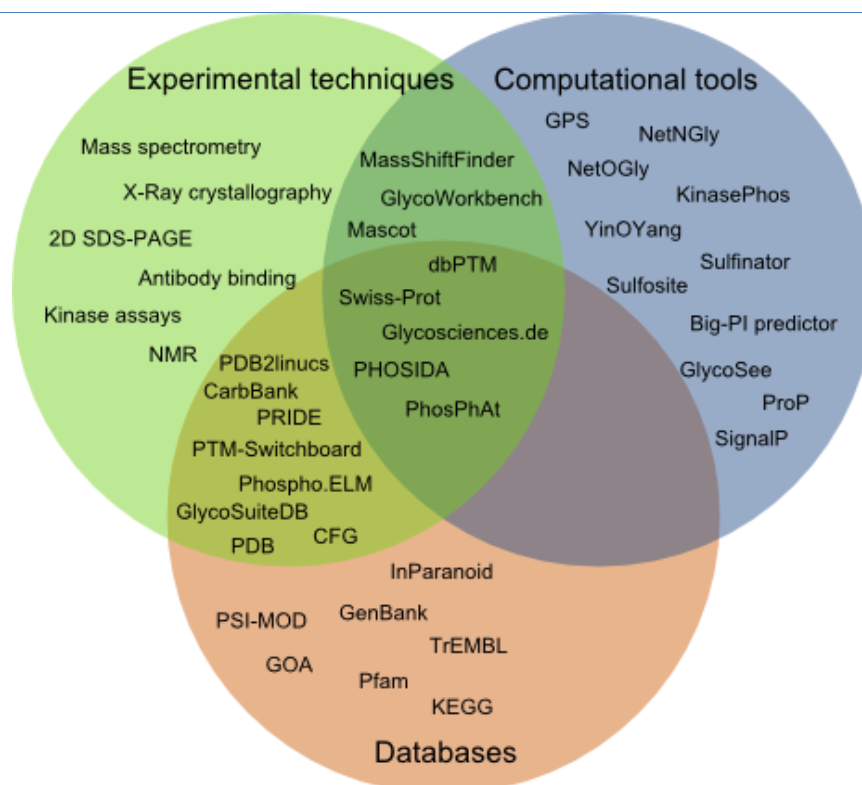


Figure 5: Intersection between PTM related tools, databases and experimental determination techniques.

Databases that store annotations for PTMs can be classified into two groups. One group stores annotations for a particular subset of modifications, whilst the other stores annotations for the majority of known modification types. Phosphorylation and glycosylation are by far the most common modifications stored in any of the current databases. These two types of modification also have a large number of databases that are specific to them.

Section 1.2.1 Phosphorylation

Phospho.ELM (<<http://phospho.elm.eu.org/>>) is a database of experimentally determined phosphorylation sites (Diella *et al.* 2004). Version 7 of this database contained 4,078 proteins with 2,083 tyrosine, 12,025 serine and 2,362 threonine modification annotations (Diella *et al.* 2008b). This database is manually curated and contains additional information such as the cellular kinase(s) responsible for modifying a particular residue (Diella *et al.* 2004). PHOSIDA (Phosphorylation Site Database) (<<http://www.phosida.de>>) is another database that contains annotations derived from mass spectrometry experiments (Gnad *et al.* 2007). In addition Gnad *et al.* (2007) designed a phosphosite predictor using a SVM (Support Vector Machine) trained on the experimentally determined sites in PHOSIDA. PhosPhAt (<<http://phosphat.mpimp-goelm.mpg.de/>>) is a database of phosphorylation sites in *Arabidopsis thaliana* determined by mass spectrometry (Heazlewood *et al.* 2008). PhosPhAt also contains a plant specific predictor of phosphorylation sites and was updated in 2010 to include predicted phosphorylation sites in the core database (Durek *et al.* 2010).

Section 1.2.2 Glycosylation

The complexity in analysing and representing glycan structures has resulted in the creation of a whole field of informatics referred to as glycoinformatics – see Frank and Schloissnig 2010 and Perez and Mulloy 2005 for a review of the field. The first database created in this field was the Complex Carbohydrate Structure Database (CCSD) created at the Complex Carbohydrate Research Centre (CCRC) (Doubet *et al.* 1989). This database was curated by experts in the field of glycobiology from around the world – containing published glycan structures and associated information (Doubet *et al.* 1989). However, funding for this database ran out in 1999 (Loss *et al.* 2002). A database of O-linked glycans was published in 1995 (Hansen *et al.* 1996), which was shortly followed by an accompanying prediction tool (Julenius *et al.* 2005). The first large scale database released after the demise of the CCSD was the GlycoSuiteDB (Cooper *et al.* 2001; Cooper *et al.* 2003). The GlycoSuiteDB was a commercial database that contained N- and O-Linked glycan structures that had been extracted from the literature (Cooper *et al.* 2001). One of the most useful

aspects of this database was their inclusion of attachment site information where it was available, as well as links to the UniProtKB (Cooper *et al.* 2001). Although useful, the commercial aspect of this database limited its uptake in the scientific community. In May 2009 the GlycoSuiteDB web interface was made public – allowing glycobiologists to access the underlying data (<<http://glycosuitedb.expasy.org/glycosuite/glycodb>>). Around the same time as GlycoSuiteDB was released, a completely free alternative the SWEET-DB was released (Loss *et al.* 2002). This database was seeded with annotations from the CCSD, NCBI PubMed, and NMR data from SugarBase (<<http://www.boc.chem.uu.nl/sugabase/sugabase.html>>). In the original publication the authors explain how they planned to expand this database with additional: structures, references, annotations and NMR spectra – using both automatic and manual curation. Lutteke, *et al.* 2004 created the PDB2LINUCS database – that contains all glycans in the PDB as well as their attachment sites (Lutteke, Frank, and von der Lieth 2004). The most recent glycodatabases have come from the CFG (Consortium for Functional Glycomics) (Raman *et al.* 2006) and EUROCarbDB initiative (<<http://eurocarbdb.org/>>). These databases focus on storing both glycan structures and the raw experimental data that were captured during experiments (for instance MS spectra) (Raman *et al.* 2006). Finally many glycoinformatics resources have been published from groups in Japan. For instance, KEGG/Glycan can be used to analyse the glycosyltransferases that are responsible for creating specific glycan structures (Hashimoto *et al.* 2006).

Section 1.2.3 UniProtKB (<http://uniprot.org>)

UniProtKB (UniProt Knowledge Base) contains by far the most PTM annotations for multiple PTM classes (Farriol-Mathis *et al.* 2004). This database is part of UniProt (Universal Protein Resource) that has been created to provide the scientific community with resources that can be used to interrogate the: structure, function and interactions of proteins (UniProt Consortium 2009). Three additional databases are included in UniProt: UniRef (UniProt Reference Clusters), UniMes (UniProt Metagenomic and Environmental Sequence database) and the UniParc (UniProt Archive) (UniProt Consortium 2009).

The UniProtKB is composed of two separate databases: Swiss-Prot and TrEMBL (Translated EMBL) (Boeckmann *et al.* 2003). The UniProtKB includes all CDSs (protein coding sequences) from: EMBL-Bank, GenBank/DBJ Nucleotide Sequence Databases, Arabidopsis Information Resource (TAIR) and Ensembl *H. sapiens* (UniProt Consortium 2009).

Swiss-Prot contains high-quality curator-maintained annotations for proteins. This includes "curator-evaluated computational analysis" (UniProt Consortium 2009). Proteins that haven't yet been curator-evaluated are placed into the TrEMBL database (Boeckmann *et al.* 2003).

All PTM annotations are curator reviewed, precluding the PTM annotation of TrEMBL entries. These PTM annotations are derived through the use of literature mining and PTM prediction tools. The PTM prediction tool chain used has not been published, although (Farriol-Mathis *et al.* 2004) states that only PTM prediction tools with low false-positive rates have been incorporated.

In Swiss-Prot release 55.3 there were 185,535 PTM annotations present on 55,445 proteins from 4,956 taxa. 276 different PTM types and 42 PTM classes were represented in this release of the Swiss-Prot database. Note that only 58,802 of these PTM annotations were not marked as being predicted.

Section 1.2.4 dbPTM (<http://dbptm.mbc.nctu.edu.tw>)

The other major database with annotations for multiple PTM classes is the dbPTM (Database of PTM) (Lee *et al.* 2006). The dbPTM aims to enhance the annotations from Swiss-Prot, Phospho.ELM, and O-GLYCBASE databases with a wide array of information (Lee *et al.* 2006). This database predominately stores the structural characteristics of modification sites, including: solvent accessibility, secondary and tertiary structure, and domain annotations (Lee *et al.* 2006). The authors of dbPTM had previously published the profile hidden markov model prediction tool called KinasePhos (Huang *et al.* 2005a). The dbPTM contains additional predictions for multiple PTM classes that were made with algorithms similar to that published for KinasePhos (Lee *et al.* 2006). Note that this database hasn't been updated since 2007.

Section 1.3 *Nomenclature*

The analysis of PTMs has been complicated by the use of synonymous and ambiguous terms (Montecchi-Palazzi *et al.* 2008). A number of attempts have been made to create controlled vocabularies (CVs), which remove such ambiguities and synonyms. Such PTM CVs are of interest to researchers in the fields of: Proteomics, Bioinformatics and Mass Spectrometry (Montecchi-Palazzi *et al.* 2008). A controlled vocabulary is simply a list of standard terms, which is usually cross-referenced to a list of synonyms for backwards compatibility. By defining relationships between terms, an ontology is formed (Stevens, Rector, and Hull 2010). Probably the best known example of an ontology created for the biological sciences, is the Gene Ontology (GO) (See Section 3.5.1(b) for further information).

Section 1.3.1 RESID

The RESID database of protein structure modifications was created in 1993 to represent the natural PTM types annotated in the Swiss-Prot and PIR (Protein Information Resource) databases (Garavelli 2003). The creation of the RESID database involved the standardisation of the terms used to describe PTMs in the Swiss-Prot database, thus creating the first PTM controlled vocabulary (Farriol-Mathis *et al.* 2004). The Swiss-Prot PTM vocabulary has however not been adopted by all researchers. Alternative vocabularies can be found in the public databases UNIMOD and DeltaMass as well as some proprietary databases (Montecchi-Palazzi *et al.* 2008).

The integration of PTM annotations has been hampered by the use of these competing vocabularies, a problem that is set to get worse with the advent of large scale proteomics initiatives (Montecchi-Palazzi *et al.* 2008). The Proteomics Standards Initiative, founded by the Human Proteome Organisation (HUPO), has started the process of creating a new controlled vocabulary of PTMs called PSI-MOD; it is hoped that this will become the standard PTM vocabulary (Montecchi-Palazzi *et al.* 2008).

The Swiss-Prot database annotates PTM events with one term from a controlled vocabulary (Farriol-Mathis *et al.* 2004). These will from here-on be

referred to as a PTM type, examples include: 4-aspartylphosphate, Phosphothreonine and O-(5'-phospho-DNA)-serine. Associative keywords are used to group together PTMs involving similar chemical changes (Farriol-Mathis *et al.* 2004). For example the three PTM terms listed previously are associated with the parent keyword Phosphoprotein.

PTM events that are derived from multiple PTM processes are associated with multiple keywords, for example N6,N6,N6-trimethyl-5-hydroxylysine is associated with the keywords Hydroxylation and Methylation. Keywords can also be parents of other keywords. Take the PTM type "GPI-anchor amidated cysteine", involving the transfer of a glycolipid onto a cysteine residue, which has the keyword "GPI-anchor". This keyword has the two parent keywords "Lipoprotein" and "Glycoprotein".

It could be argued that by linking PTM types to PTM classes an 'is_a' relationship is being formed; creating a Swiss-Prot PTM ontology (although it is never referred to as such).

Section 1.3.2 PSI-MOD

The Proteomics Standards Initiative has initiated the construction of an ontology of PTMs. They decided that this was necessary as none of the vocabularies in current use fitted the needs of the Proteomic Informatics (PSI-PI) and Molecular Interactions (PSI-MI) working groups. The PSI aims for PSI-MOD to become the standard PTM vocabulary both in HUGO and in the wider scientific community. (Montecchi-Palazzi *et al.* 2008).

An ontology is the specification of "an abstract, or simplified view of the world that we wish to represent for some purpose" (Gruber 1993). Ontologies contain terms or classes that represent abstract views of real world objects; as well as containing a representation of the relationships between such objects (Gruber 1993). The Gene Ontology (Ashburner *et al.* 2000) is the canonical ontology in the bioinformatics community. The gene ontology contains three namespaces that allow for scientists to annotate the molecular function, biological process, and cellular component of gene products (Ashburner *et al.* 2000). The gene ontology is composed of a DAG (Directed Acyclic Graph) where terms are the

nodes, and edges represent the relationships between terms (Ashburner *et al.* 2000). The presence of an 'is_a' relationship between two terms indicates that one term is a subtype of the other (Smith *et al.* 2005). The 'part_of' relationship is used to indicate that a term is contained in the object represented by the other/parent term (Smith *et al.* 2005). The lack of correctly defined and conceived relationship-type definitions has been a particular problem in the bioinformatics community and is addressed in detail in Smith *et al.* 2005.

PSI-MOD uses the four relationship types as shown in Table 2. Examples of each of these types can be seen in Table 3.

Relationship	Definition
is_a	Indicating membership of a class
part_of	Indicating a child entity is a substructure
derives_from	Indicating a child entity is the result of a chemical process
has_function_parent	Indicating that a child entity derived by functional modification shares at least one characteristic group with the parent entity.

Table 2: PSI-MOD relationships – taken from (Montecchi-Palazzi *et al.* 2008).

Child Term	Relationship	Parent Term
O-glycosyl-L-threonine	is_a	O-glycosylated residue
pentosylated residue	part_of	complex glycosylation
N6,N6,N6-trimethyl-L-lysine (from L-lysine residue)	derives_from	protonated L-lysine (L-lysine) residue
neddylated lysine	Contains	N6-glycyl-L-lysine

Table 3: PSI-MOD relationship examples.

Section 1.4 Aims of thesis

The main goal of this body of work was to create the tools and resources necessary to allow for the analysis of PTM conservation. Two different conservation analyses were considered to be important at the start of this work. The first use-case involves users who wish to analyse PTM conservation for a specific list of proteins. The second use-case involves users who wish to analyse PTM conservation between taxonomic groups; for example a user may wish to know what percentage of the *H. sapiens* phosphoproteome is conserved in *E. coli*. The primary purpose of the majority of the databases previously described is to act as a store of experimentally-determined, or computationally-predicted, modification annotations. Few provide the interrogation interfaces that are required by experimentalists to perform even the simplest of conservation analysis. Of course there are always exceptions and the

PHOSIDA is probably the best example. This database allows users to identify conserved phosphorylated residues between orthologous sequences. However there is currently no database that allows for a simple conservation analysis to be performed for many different types of PTM classes. Section 5.3 contains a more detailed description of existing resources and how they can be used in workflows to analyse PTM conservation.

Software that is going to perform a conservation analysis for scientists requires access to many different pieces of information. For example to analyse the conservation of the phosphoproteome between *Eukaryotes* and *Mammals* requires access to: taxonomic trees, proteome sets, PTM annotations and some form of homology annotation. The following three chapters of this thesis describe the integration of the many different resources required by the tool that has been designed to analysis PTM conservation.

Chapter 2 starts by describing the creation of the PTMDB, a database that integrates PTM annotations from Swiss-Prot, Phospho.ELM and the PDB. This chapter highlights the difficulties in analysing the conservation of glycosylation caused by incomplete structural datasets.

Chapter 3 describes the process of adding orthology assignments to the PTMDB. Incorporating these annotations into the PTMDB allows PTM Browser to perform a conservation analysis on a choosen protein – without requiring users to manually specify proteins they wish to compare (although they still can). In addition these orthology assignments can be used by PTM Browser to carry out a high-throughput analysis of the conservation of modifications between orthologues of user selected species.

Chapter 4 describes a process that has been used to transfer existing PTM annotations onto homologous sites. Homologous sites are identified through Pfam domain alignments that have been incorporated into the PTMDB. As well as providing the scaffolding for this process, the domain annotations also allow for a domain level conservation analysis to be performed by PTM Browser. The cross-annotation procedure combined with the Pfam alignment integration, allows PTM Browser to quickly analyse PTM conservation without having to carry out a time-consuming multiple sequence alignment step. The majority of

the PTM annotations in the PTMDB were imported from Swiss-Prot. Therefore species whose proteome is predominantly located in the TrEMBL portion of the UniProtKB are underrepresented in the PTMDB. This cross-annotation procedure addresses this limitation to an extent by including proteins from the TrEMBL database. Therefore the PTMDB contains the first large scale PTM annotation for some species whose genomes have recently been sequenced.

Chapter 5 describes the development of the PTM Browser tool which allows users to analyse the conservation of PTMs. Included in this chapter is a description of a web service that has been created to allow programmatic access to both the PTMDB and the PTM Browser conservation analysis routines.

Chapter 6 describes the results of a number of different PTM conservation analysis performed with the PTM Browser tool. This includes an analysis on the conservation of modifications between *H. sapiens* and a number of other species. An example of PTM Browser being used to analysis a protein family is presented, via an analysis of the conservation of modifications in the p53 family of proteins.

Chapter 2

A new database of PTMs

Section 2.1 *Summary*

There are many post-translational modification databases in existence that store a variety of information. These databases, and the tools that accompany them, are predominately focused on allowing users to retrieve a list of known modification sites for a specific protein. Some databases, like UniProtKB, also provide limited information on the conservation of modifications across homologous proteins. However, using existing resources, it is currently impossible for users to ask simple questions such as “What modification sites appear in human proteins but not in their mouse counterparts?”. Before software can be designed to answer such questions it is necessary to decide where this software is going to obtain its annotations. None of the currently available databases are designed to be used in such a fashion. Therefore presented in this chapter is a description of the creation of a new database of PTMs. This database has been designed with the specific aim of allowing such questions to be answered. The seeding of this database with annotations from UniProtKB, Phospho.ELM and the PDB is described. The extension of this database to include additional annotations, and other information, is then described in the following two chapters.

Section 2.2 *Software and conventions*

Throughout this and the following chapters there will be references to many classes and scripts that carry out particular tasks required for the work presented in this thesis. The first time a class name is provided it will be written in fully qualified form, e.g. <org.drd20.util.MySqlUtil>. From then on all references to the class will appear unqualified and underlined, e.g. <MySqlUtil>. Unless otherwise stated all classes reside in the Perl project Bioinformatics Base Perl (BBP), information on obtaining and installing the software described in this thesis can be found in Appendix 2. All of the classes described are designed for reuse in other pieces of software. However some include a static main method implementation similar to that of Java classes, which allows the class to be called directly by a Perl interpreter that runs some default task. When a particular task being described is accomplished by running the static main method of a class its name will be suffixed with a question mark.

To make it clear that a table, namespace or field from a relational database is being referred to they will be underlined, presented in bold font, and enclosed in angle brackets, e.g. <**vocabulary.PtmDescription**>.

Section 2.3 *Design approach*

This section outlines aspects of the schema design process, which might be unfamiliar or require justification to the reader.

The MySQL (<<http://mysql.com>>) RDBMS (Relational Database Management System) was chosen to store the PTMDB primarily because of its ease of use and support. Unless stated otherwise all tables were designed to be stored using the InnoDB storage engine (at the time of development the default for MySQL was MyISAM, which doesn't support foreign key constraints). PostgreSQL (<<http://www.postgresql.org>>) users should be aware that MySQL doesn't support the notation of separate namespaces in the same database. It's common practise in the MySQL community to represent each namespace with its own separate database, this methodology has been adapted for the schemas of the PTMDB.

The majority of schemas were designed using the MySQL Workbench (MWB) tool (<<http://wb.mysql.com/>>). Entity relationship diagrams were produced using the SVG (Support Vector Graphic) export functionality of MWB. Note these images have been manually adjusted (using Inkscape <<http://inkscape.org>>) to enable primary, surrogate and business keys to be distinguished. MWB determines the relationship type (identifying or non-identifying) between two entities based on the primary key of the child table. Where a surrogate primary key is used the relationship types determined by MWB aren't particularly informative. Therefore the relationship types exported by MWB for entities with a surrogate primary key have been manually adjusted to reflect the business key. An identifying relationship indicates that a row in a child table cannot be identified without the foreign key constraint to the parent table. In contrast a non-identifying relationship indicates that a row in a child table can be identified without the foreign key constraint to the parent table - refer to the caption of Figure 10 for examples of each.

The primary key of a table includes those fields without which an entity cannot be uniquely identified. A surrogate primary key is used where a suitable natural primary key doesn't exist or it's advantageous not to use it. For instance surrogate primary keys are commonly used by ORM (Object Relational Mapping) technologies (e.g. Hibernate (<<http://www.hibernate.org>>) as a fast way of uniquely identifying an entity. Surrogate primary keys have been used wherever possible in the PTMDB for both ease of schema design and future compatibility with ORM technologies. Inserting a row into a table that has a foreign key dependency on another table's surrogate primary key can be problematic as it requires an additional query to discover the correct surrogate primary key to insert. To resolve this potential issue a new Perl class <org.drd20.util.AbstractERD.WorkBenchParser> has been created that is able to create an object model of the schema using SQL that MWB is able to export. This object model allows for data to be inserted into a table without having to lookup the surrogate primary key manually. This software has to be able to workout which fields represent the business key of a table in order to successfully resolve the surrogate primary key. A business key like a natural primary key represents those fields without which an entity cannot be identified.

For each table that requires a business key to be defined, a unique index called Entry is created, that includes the fields that identify an entity in the corresponding table. These unique indexes can be added to the schema using MWB, the `<WorkBenchParser>` class has been designed to extract the business key from unique indexes called Entry.

Entity relationship diagrams are shown throughout this chapter, refer to Figure 6 to see what relationships exist between them.

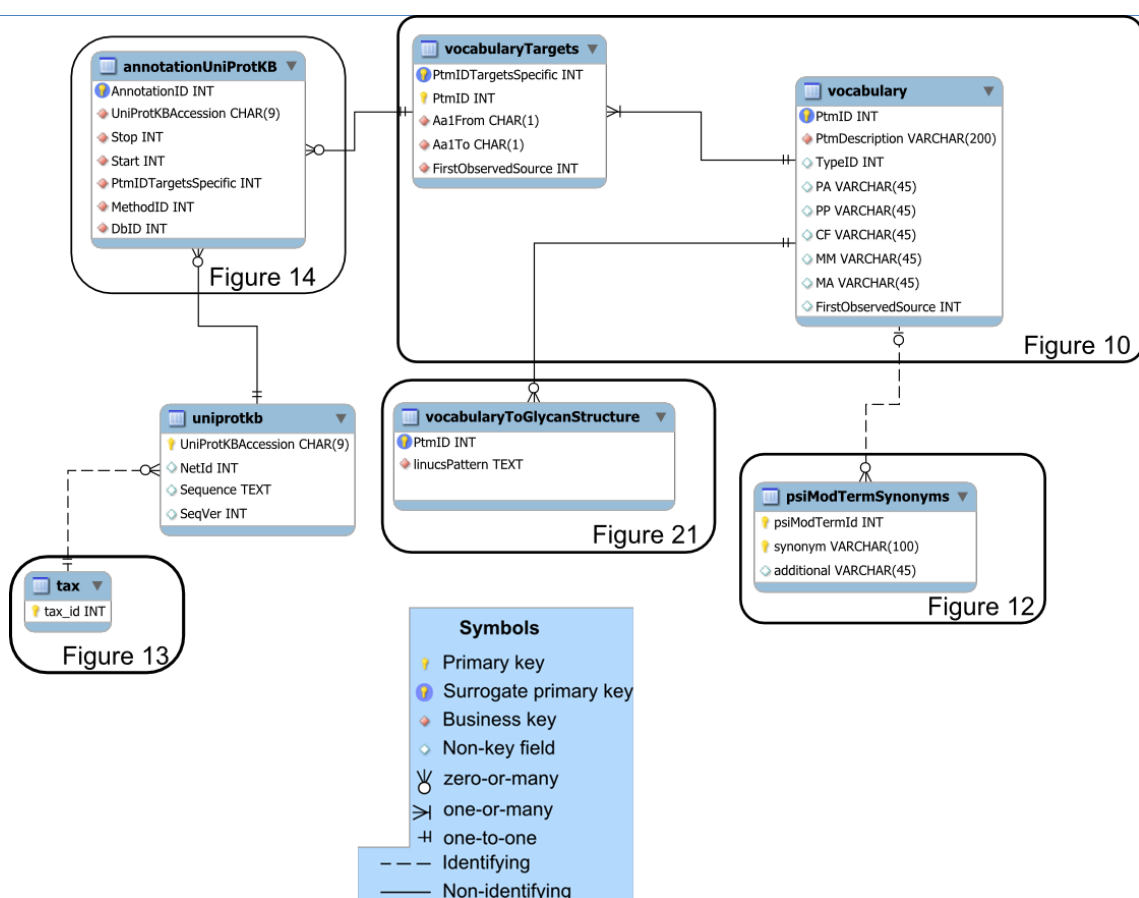


Figure 6: Entity relationship figure connections. Figure displays the relationships between the different ERD figures shown throughout this chapter.

Section 2.4 Vocabulary

At the time of writing two popular nomenclatures were being actively developed for the purpose of describing PTMs. The first to be developed was designed for use in the Swiss-Prot database, for the purpose of unambiguously annotating PTM events (Farriol-Mathis *et al.* 2004). The second was published 4 years later by the Proteomics Standards Initiative (PSI) for use by the Proteomic Informatics (PSI-PI) and Molecular Interactions (PSI-MI) working groups

(Montecchi-Palazzi *et al.* 2008). The PSI designed their nomenclature as an ontology, which they called PSI-MOD. Further details on PTM nomenclatures and the differences between them can be found in Section 1.3.

In Section 1.2 an overview of existing PTM databases was presented, which clearly demonstrated Swiss-Prot as having the most diverse and numerous PTM annotation dataset. The new PTM database that is being described, the PTMDB, incorporates the majority of its annotations from the Swiss-Prot database. For this reason the primary PTM nomenclature used in the PTMDB has been derived from that which is used in the Swiss-Prot database. In the future it is likely that database curators will annotate PTM events with purposely designed ontologies, as has already happened for describing gene products and their functions using the Gene Ontology (Ashburner *et al.* 2000). Therefore support for the PSI-MOD ontology has also been added to the PTMDB.

The following two sections describe the implementation of both these nomenclatures in the PTMDB.

Section 2.4.1 Swiss-Prot vocabulary

The following description of the Swiss-Prot PTM vocabulary is based on information contained in the original paper by Farriol-Mathis *et al.*, (2004) and the UniProtKB manual (<<http://www.expasy.ch/sprot/userman.html>>). This nomenclature is referred to by the authors as a controlled vocabulary, downloadable from the following URL: <<http://www.uniprot.org/docs/ptmlist>>. There are three levels of classification in this vocabulary: (i) Feature key, (ii) Feature description and (iii) Keyword. The most specific of these is the feature description; examples include Phosphoserine, N-acetylserine, N-acetylmethionine, N6-acetyllysine, etc. The name of this level of classification isn't particularly descriptive for the purposes of a PTM vocabulary, therefore the PTMDB feature description has been renamed to PTM type. The second layer of this vocabulary utilises UniProtKB keywords to group similar PTM types. For example Phosphoserine, Phosphothreonine, Phosphotyrosine, and Phosphohistidine all belong to the keyword Phosphoprotein. For convenience these keywords are referred to in the PTMDB as PTM classes. Note that the file that contains this vocabulary also contains an index between PTM types and

their associated PTM classes. The top level of this vocabulary groups PTM types by six feature keys, which are shown in Table 4.

Feature Key	Description
LIPID	PTM types with a predominantly lipid constitution.
CROSSLNK	Formation of a covalent bond between two amino acids.
DISULFIDE	Disulphide bonds.
CARBOHYD	PTM types with a predominantly glycan component.
CHAIN	Peptide chain cleavage annotations.
MOD_RES	All PTMs not covered by the previous feature keys.

Table 4: UniProtKB feature keys associated with PTM types.

An idealised version of the Swiss-Prot PTM vocabulary schema is shown in Figure 7, in summary the figure demonstrates the following points:

1. All feature keys listed in Table 1 are associated with one-or-more PTM types.
2. PTM types are associated with zero-or-more PTM classes.
3. Every PTM class is associated with one-or-more PTM types.
4. All PTM types are associated with one-or-more PTM classes.
5. A PTM class can have zero-or-more parents.
6. A PTM class can have zero-or-more children.

In order to understand points 4-6 it's important to understand the connections that can exist between UniProtKB keywords. It's already been stated that a PTM class is a UniProtKB keyword; such keywords can be connected to others in parent-child fashion. Note that keyword definitions and the connections that exist between them can be downloaded from the following URL: <http://www.uniprot.org/docs/keywlist>. An excellent example of this is provided by the GPI-anchor keyword, which has the parents Lipoprotein and Glycoprotein; remember that this type of modification includes both a lipid and carbohydrate component. Note that the PTM type-to-PTM class index, provided with this vocabulary, doesn't explicitly map to parent keywords; for example the PTM type GPI-anchor amidated alanine is only associated with the keyword GPI-anchor. Some modifications are the result of two separate processes for these multiple keywords are listed in the provided PTM type-to-PTM class index. For example the PTM type 5-hydroxy-3-methylproline is mapped to the PTM classes Hydroxylation and Methylation in this mapping.

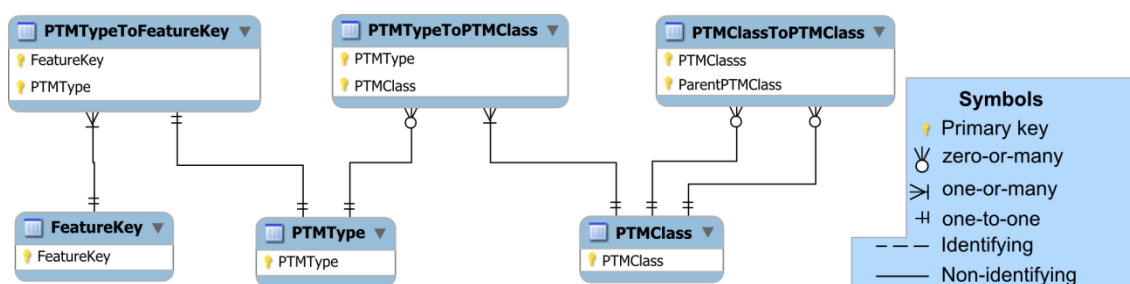


Figure 7: An idealised version of the Swiss-Prot PTM vocabulary schema. This diagram was designed using MySQL Workbench (<http://wb.mysql.com/>), and exported as an SVG (Support Vector Graphic). Note that entity names, mapping table names, and field names do not directly correlate with the terminology used in either Farriol-Mathis et al., (2004) or the UniProtKB manual.

Table 5 contains a list of the additional information that this vocabulary stores for each PTM type. Although this data isn't pertinent to the task of analysing the conservation of modifications (the aim of this body of work), for the sake of completeness it has been included in the schema that will be discussed below.

Short identifier	Plain Text Description	Occurrence
ID	Identifier (FT description)	Once
AC	Accession (PTM-xxxx)	Once
FT	Feature Key	Once
TG	Target	Once
PA	Position of the modification on the amino acid	Optional, once
PP	Position of the modification in the polypeptide	Optional, once.
CF	Correction formula	Optional, once
MM	Monoisotopic mass difference	Optional, once
MA	Average mass difference	Optional, once
LC	Cellular location	Optional, one or more
TR	Taxonomic range	Optional, one or more
KW	Keyword	Optional, one or more
DR	Cross-reference to PTM databases	Optional, one or more

Table 5: UniProtKB PTM vocabulary attributes – list adapted from <http://www.uniprot.org/docs/ptmlist>.

Section 2.4.2 Primary PTMDB vocabulary

For reasons previously explained it was decided that the primary PTM vocabulary of the PTMDB should be based on that used in the UniProtKB. What follows is a brief description of the schema that has been designed to house this vocabulary. Throughout this description the reader is encouraged to refer to Figure 10, which contains a schema diagram specific to the vocabulary section of the PTMDB. At the heart of the schema is the **<vocabulary>** table that contains a list of PTM types and their associated feature key. In addition this table also includes all attributes from Table 2, which can occur at most once. The remaining one-or-many attributes from Table 2 have been mapped on the tables: **<vocabularyLocalisation>**, **<vocabularyTaxonomicRange>**,

and **<vocabularyKeywords>**. Not all the PTM types and classes stored in the primary PTMDB vocabulary will have been imported from the Swiss-Prot PTM vocabulary. For this reason the **<vocabulary>** table also includes the field **<vocabulary.FirstObservedSource>** to enable a PTM type to be associated with the database it was first imported from.

The terms used to describe PTM types in the Swiss-Prot PTM vocabulary are not all amino acid specific. This shouldn't come as a particular surprise as it would make the vocabulary incredibly redundant; especially when you consider the number of glycan structure/amino acid pairs that could theoretically exist (KEGG/Glycan (Hashimoto *et al.* 2006) currently contains 11,000 structures). Early on in the database design process it was decided that it would be useful to be able to extract PTM annotations with matching PTM types and attachment amino acids. To meet this requirement PTM annotations are stored in the PTMDB using a PTM type identifier that is amino acid specific. The table **<vocabularyTargetsSpecific>** contains these new identifiers associated with their corresponding PTM type and amino acid. PTM types that are clustered under the feature keys DISULPHIDE and CROSSLNK involve multiple amino acids; it's for this reason that the table **<vocabularyTargetsSpecific>** contains two amino acid fields for all other PTM types both fields contain identical values.

The Swiss-Prot vocabulary flat file that was mentioned earlier doesn't include any of the Glycosylation PTM types found in the Swiss-Prot database. Glycan structures are by far the most structurally complex of all PTMs, exemplified by the numerous encoding languages that have been created (Glyde I (Hellen *et al.* 2008) (Sahoo *et al.* 2005), Glyde II (York 2011), GlycoCT (Herget *et al.* 2008), LINUCS (Bohne-Lang *et al.* 2001), etc.). In Swiss-Prot Glycosylation PTM types use a simple encoding scheme that indicates the linkage type, terminal reducing-end sugar, and an indicator for whether any sugars are attached to the terminal reducing-end sugar. See Figure 8 for further details and Glycosylation PTM type examples.

REGEXP[OCNS]{1}-linked \ (SUGAR(?:...)?(?: or SUGAR(?:...)?)?\)
SUGAR=(?:HexNAc|GlcNAc|GalNAc|Man|Glc|Gal|Hex|Xyl|Ara|Fuc){1}

(a)

N-linked Glycosylation PTM types

N-linked (GlcNAc...)

N-linked (Glc...)

N-linked (GlcNAc)

N-linked (GlcNAc or GlcNAc...)

N-linked (Man)

(b)

Figure 8: Swiss-Prot Glycosylation PTM type format. (a) Perl regular expression which conforms to the Swiss-Prot Glycosylation PTM type format (documentation regarding the regular expression notation can be found at the following URL <<http://perldoc.perl.org/perlre.html>>). (b) All of the N-linked Glycosylation PTM types in Swiss-Prot.

The PTMDB vocabulary tables are populated with the Swiss-Prot PTM vocabulary using the script <ptmDB/Main/PTMKeywords/ptmVocabularyErd.pl>. As the Glycosylation PTM types aren't included in the Swiss-Prot PTM vocabulary flat file they have to be imported dynamically as PTM annotations are parsed from the Swiss-Prot database. Additional keywords have also had to be created for the Glycosylation PTM types. Its common practise to discuss Glycosylation based on the linkage between the protein and the glycan, therefore the following new keywords have been created: Glycosylation, Glycosylation_O_Linked, Glycosylation_N_Linked, Glycosylation_C_Linked, and Glycosylation_S_Linked. Note that the Swiss-Prot PTM vocabulary flat file maps PTM types to the amino acids that they can occur on, therefore for all PTM types represented in this file the table <**vocabularyTargetsSpecific**> can be populated. For Glycosylation PTM types this table is dynamically populated as Swiss-Prot PTM annotations are parsed. Note that in the PTMDB vocabulary all PTM types that don't have a PTM class in the Swiss-Prot PTM vocabulary have been associated with the new PTM class – Other.

Table 4 contains the six feature keys which represent all PTM types that can be stored in the Swiss-Prot database. Note that it was decided that although the PTMDB vocabulary and annotation schemas would be designed to support all six; the following three are currently ignored by both the vocabulary import and annotation import systems – CROSSLNK, DISULFIDE, and CHAIN.

Figure 9 displays the number of PTM types that are grouped under each PTM class in the PTMDB vocabulary. In addition this figure shows the number of additional identifiers which had to be generated to create amino acid specific PTM types. Note that this figure was generated after the import of Swiss-Prot PTM annotations so that data could be shown for the Glycosylation PTM classes. This figure shows that there are a great many PTM types that have no associated PTM class in the Swiss-Prot PTM vocabulary. A total of 288 PTM types grouped by 40 PTM classes are in the current PTMDB vocabulary.

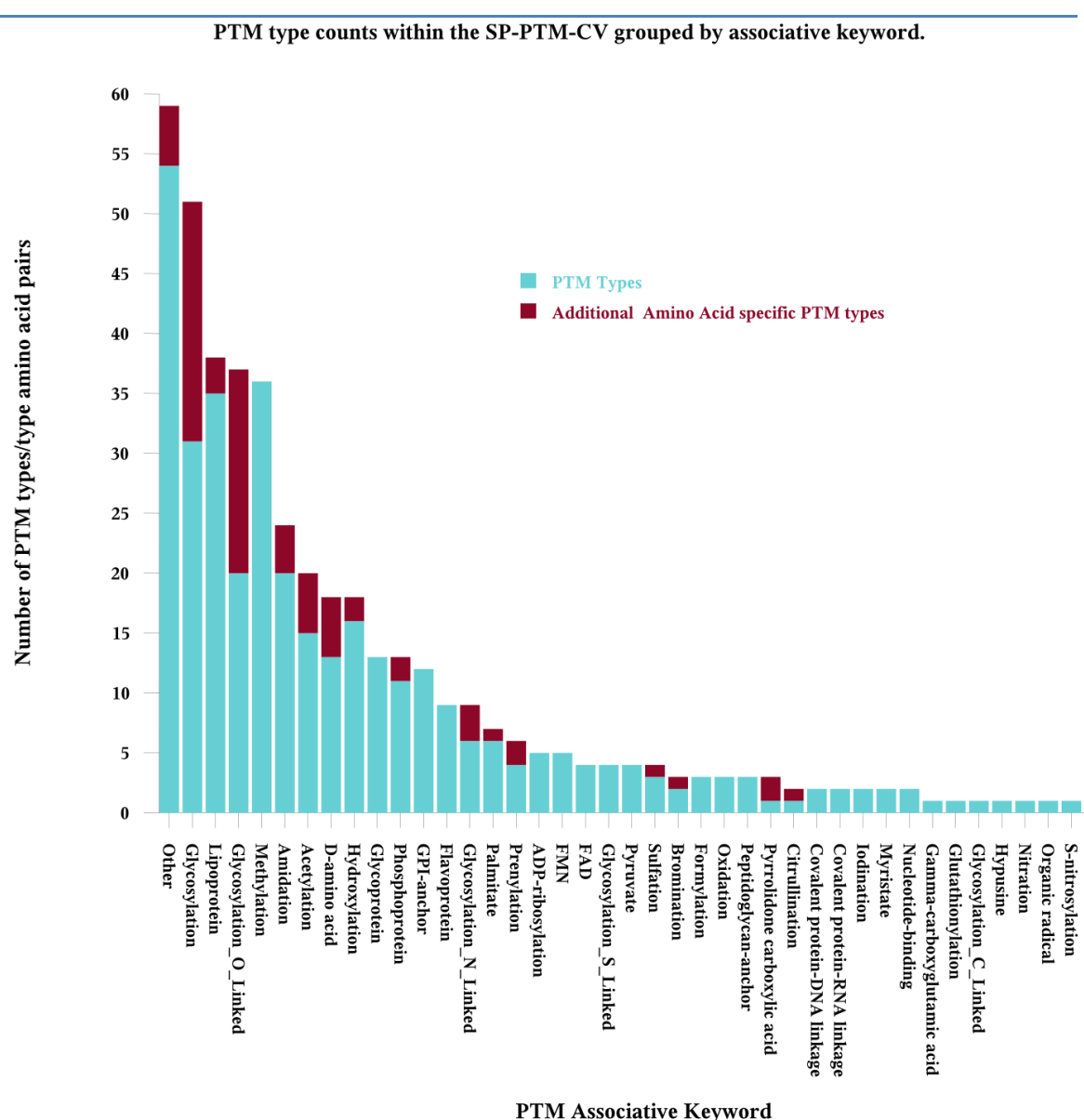


Figure 9: Distribution of PTM types by PTM class in the PTMDB vocabulary. The number of additional amino acid specific PTM types created represents the difference between the number of corresponding rows (for a given PTM class) on the tables [vocabulary](#) and [vocabularyTargetsSpecific](#). This figure was generated after Swiss-Prot PTM annotations had been imported into the PTMDB so that the Glycosylation PTM classes could be included.

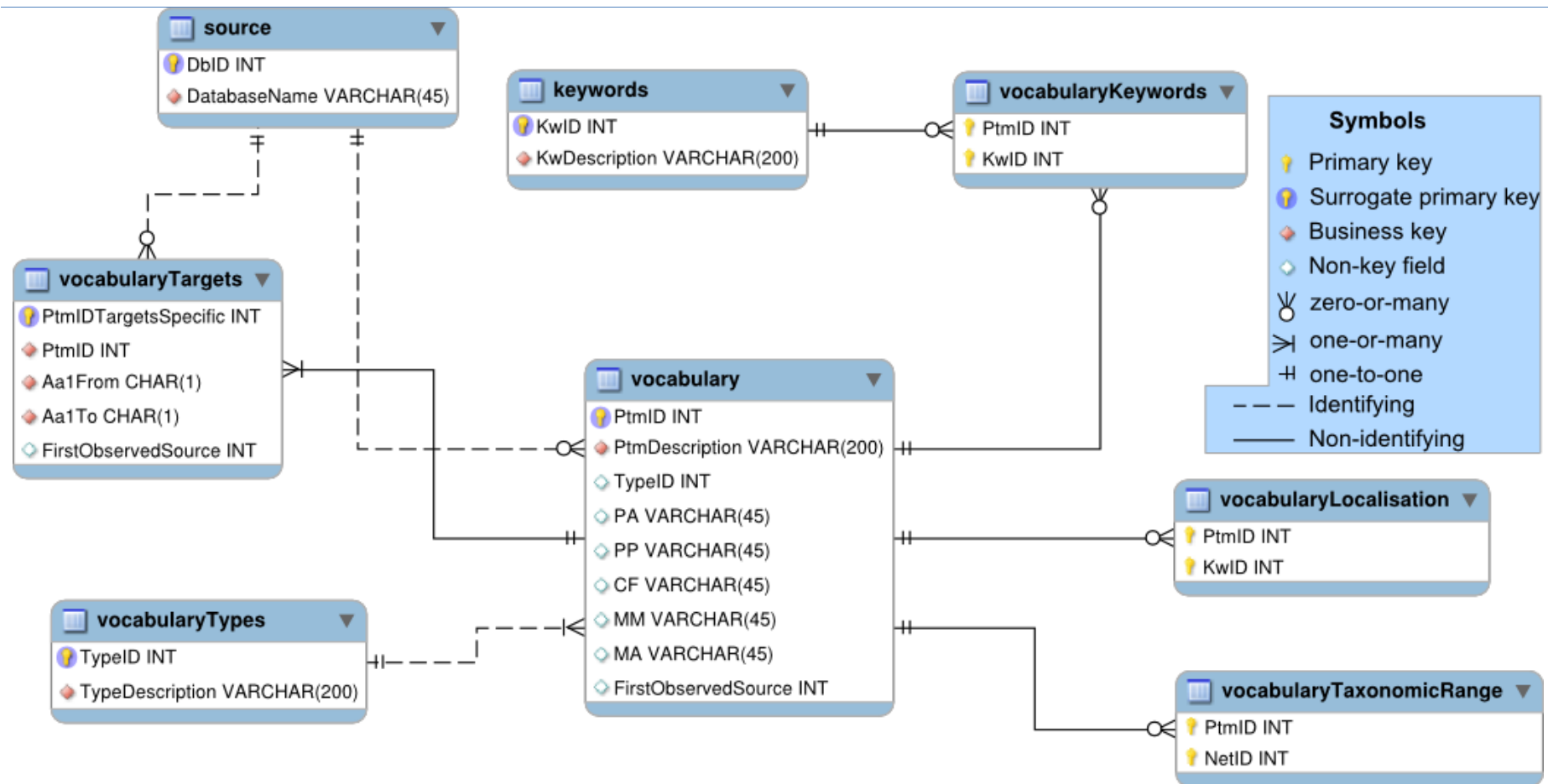


Figure 10: Post-translational modification vocabulary schema diagram. *Identifying and non-identifying relationships:* a) field PtmID is included in the business key of vocabularyTargets so the relationship between vocabularyTargets and vocabulary has been set to identifying, b) TypeID is not included in the business key of vocabulary so the relationship between vocabulary and vocabularyType has been set to non-identifying. The tables in this figure have been placed in the PTMDB namespace.

Section 2.4.3 PSI-MOD incorporation

PSI-MOD support has been added to the PTMDB for two reasons: a) it is likely to become the PTM controlled vocabulary standard, and b) the relationships represented in the graph structure will allow for more powerful queries to be constructed than are possible with the Swiss-Prot PTM vocabulary. PSI-MOD contains all of the entries from the RESID, Delta Mass and UniMod databases (Montecchi-Palazzi *et al.* 2008). As stated previously the RESID database contains all the PTM types found in the Swiss-Prot PTM vocabulary and hence the PTMDB vocabulary, therefore it's possible to map PTM types in the PTMDB to their PSI-MOD equivalents.

The PSI-MOD ontology is supported in the PTMDB via the schema shown in Figure 12. Terms in the PSI-MOD have a number of additional attributes associated with them; those that can be imported into the PTMDB are shown in Table 6. Note that not all of the terms have all of these attributes. The table **<psiModTerms>** is used to store a list of all PSI-MOD terms and these additional attributes associated with them.

Short Name	Description
Formula	Chemical formula
MassAvg	Monoisotopic mass
DiffFormula	Formula for the difference between observed modified residues and an originating residue
DiffAvg	Chemical average mass difference
DiffMono	Monoisotopic mass difference
Source	Residue source, currently with values of 'Natural', 'Artifact' or 'Hypothetical'
Origin	Residue origin with a value of one IUPAC standard single-letter amino acid code or for cross-links, one for each participating residue.
TermSpec	Protein amino or carboxyl terminus specificity, currently with a value of 'C-term' or 'N-term' or 'none'.

Table 6: PSI-MOD term attributes supported by the PTMDB. List taken directly from the original specification of PSI-MOD found in Montecchi-Palazzi, 2008.

The PSI-MOD schema shown in Figure 12 has the following characteristics.

1. A term may have zero-or-more synonyms.
2. A term may represent the modification of zero-or-more amino acids.
3. A term may have zero-or-more parents and zero-or-more children.
4. A term may be a part of zero-or-more subsets (a slim).

The synonyms of a PSI-MOD term represent the equivalent terms in the RESID, Delta Mass and UniMod databases. PSI-MOD terms that are equivalent to PTMDB PTM types can be identified by matching values in the fields **<vocabulary.PtmDescription>** and **<psiModTermSynonyms.synonym>**. The reason not all PSI-MOD terms are associated with target amino acids, results from their abstract nature; for example MOD:00001 is the parent of all terms which represent the alkylation of residues. Note that all PSI-MOD terms have a parent term except for the root term of the ontology (MOD:00000 protein modification). As has been explained previously a leaf term is one that doesn't have any children. The majority of the PTM types in the PTMDB vocabulary map to leaf terms. The PSI-MOD graph is stored in the table **<psiModTermRelationships>**. To allow for all direct and indirect descendants of a specific PSI-MOD term to be obtained, without reparsing the PSI-MOD graph, the table **<psiModTermToAllParents>** has been created.

The PSI-MOD ontology is available in OBO (Open Biological and Biomedical Ontologies) format from the following URL **<<http://psidev.sourceforge.net/mod/data/PSI-MOD.obo>>**. The shell script **<org.drd20.bioinformatics.database.ptmdb.psiModAutomate.sh>** (part of the BBP project) has been designed to read the PSI-MOD OBO flat file and import data into the PTMDB PSI-MOD schema. Note that this script makes use of a number of Perl and Java programs to accomplish this task. The revision of the PSI-MOD ontology incorporated into the PTMDB being reported here contained 1,409 terms.

The PSI-MOD ontology contains four main paths that allow for terms to be found according to chemical process, target amino acid, isobaric sets and uncategorised modifications (Montecchi-Palazzi *et al.* 2008). Almost all of the direct descendants of the term chemical process have an equivalent PTMDB PTM class.

Figure 9 displays the number of PTM types grouped by PTM class for the PTMDB primary vocabulary. These numbers can be compared with those in Figure 11, which shows the number of terms below each direct descendant of the chemical process term.

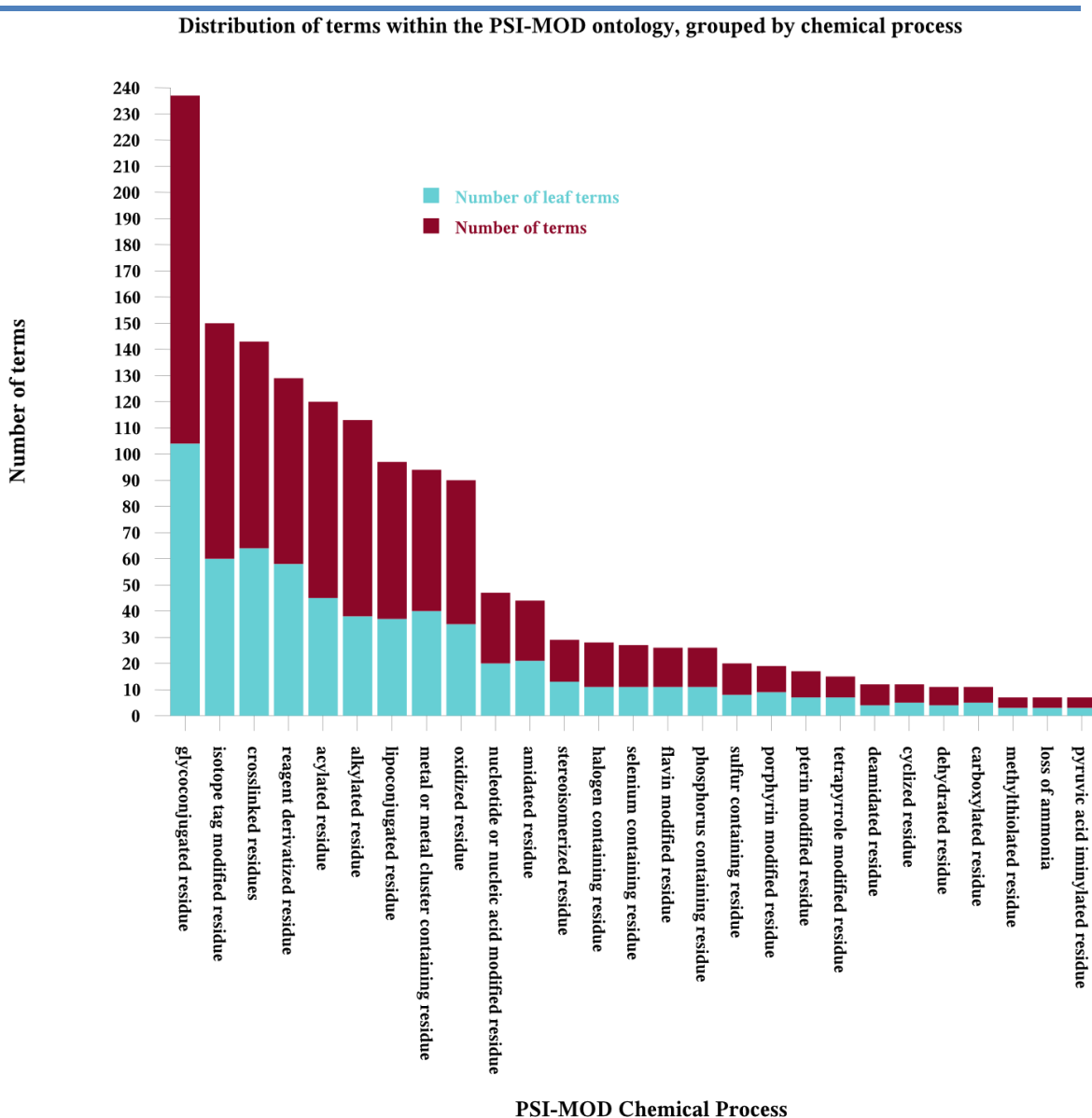


Figure 11: PTM type distribution in the PSI-MOD ontology. The x-axis contains all those terms that are direct descendants of the [chemical process](#) term (MOD:01157). This graph shows both the number of terms that can be found below each term listed on the x-axis; additionally the number of these terms that are leaf terms is also shown.

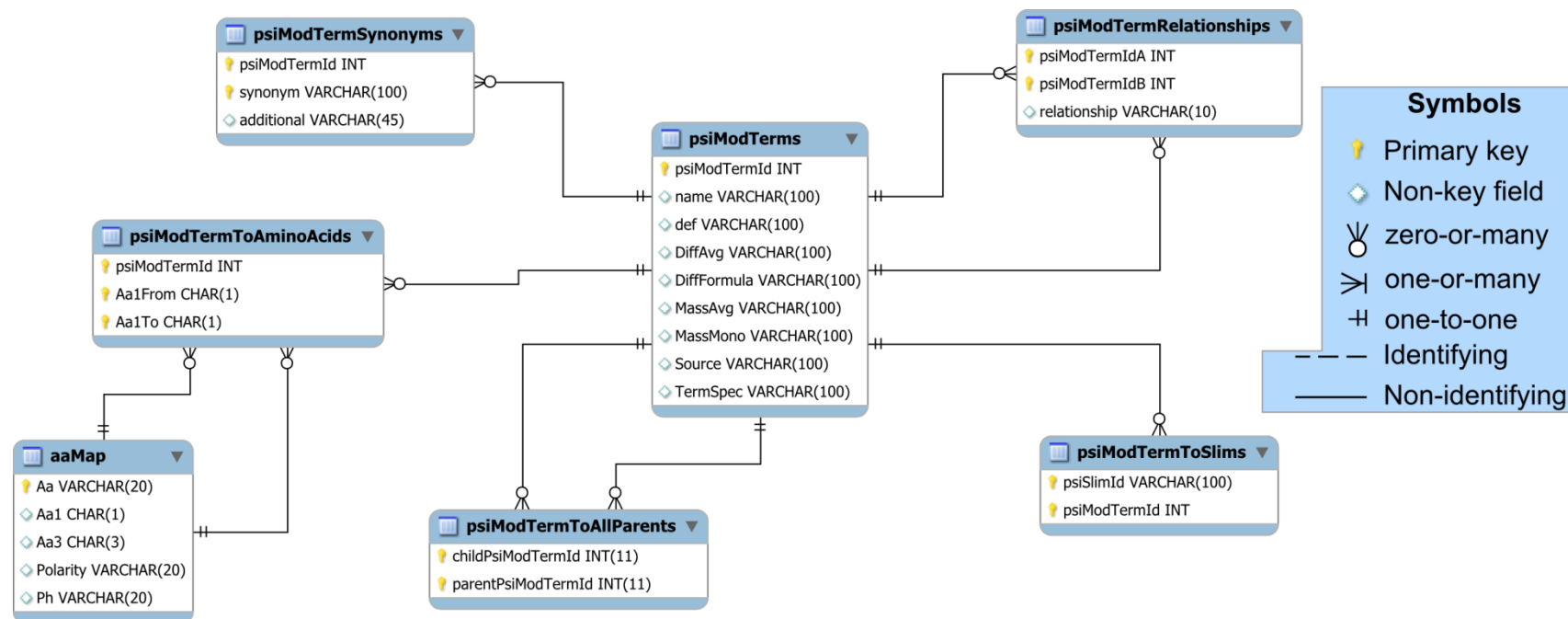


Figure 12: PSI-MOD ontology schema diagram.

Section 2.5 Annotation

Section 2.5.1 Swiss-Prot annotation format

The following description of the Swiss-Prot PTM annotation format is based on information contained in the original paper by Farriol-Mathis *et al.*, (2004) and the UniProtKB manual (<<http://www.expasy.ch/sprot/userman.html>>). A more general introduction to the Swiss-Prot database and UniProtKB can be found in Section 1.2.3.

In Swiss-Prot each protein is represented as a single entry in the database. Each entry is composed of a set of fields that describe the functional and structural attributes of the protein. The feature table component of an entry stores the annotations for splice variants, variants, signal sequences, PTMs, and many other attributes. A single annotation is represented as one line in the feature table. The type of annotation present on a line is denoted by a feature key. Feature keys were discussed previously (Section 2.4.1), six of which are specific to the PTMs listed in Table 4.

The vast majority of PTM types are associated with the MOD_RES feature key. Example PTM annotations taken from the feature tables of Swiss-Prot entries are shown in Table 7.

	Feature key	Start	Stop	Feature description
FT	MOD_RES	247	247	Phosphothreonine; by PKB (By similarity)
FT	MOD_RES	2	2	N-acetylserine (By similarity)
FT	MOD_RES	52	52	N6-acetyllysine (By similarity)
FT	MOD_RES	1	1	N-acetylmethionine

Table 7: Swiss-Prot PTM feature table example entries.

A PTM feature table entry always includes a feature key, start residue, end residue, and a feature description. As previously described (Section 2.4.1) for the majority of PTM types the start and end residue will be the same, the exceptions are those PTM types that involve multiple residues e.g. cross-links. A feature description always includes a PTM type and an evidence qualifier. The evidence qualifier may be followed by variant and/or splice variant identifiers, and finally identifiers to indicate enzymes that are responsible. Variant or splice-variant identifiers are used to indicate that the PTM being

described has an observed sequence that differs from the main sequence listed for the entry.

Swiss-Prot uses four distinct evidence tags for PTM annotations that reflect the degree to which experimental evidence supports an annotation. The tag with the highest confidence associated with it is Experimental. However note that this qualifier is never actually used, instead the absence of the other three implies that the evidence qualifier is experimental. An annotation must be both published and confirmed by experimental means to receive such an annotation.

Where the experimental evidence supporting the PTM is inconclusive the tag: Probable is used instead.

During the curation process of a Swiss-Prot entry, a PTM annotation may be cross-annotated to it from another entry. The entry must be of a homologue from a sufficiently related species; note that this rule is not qualified any further in the UniProtKB manual. There must also be evidence that suggests that the protein will be present at the right cellular location to receive the PTM. PTM cross-annotations receive the tag By similarity.

PTM annotations that are derived from the prediction tool chain used in Swiss-Prot are given the tag Potential. Unfortunately neither the prediction program nor p-value (or equivalent) is published beside such annotations.

Section 2.5.2 PTMDB annotation schema

The PTMDB annotation storage schema which is shown in Figure 14 has been designed to incorporate all of the information that can be stored in a Swiss-Prot PTM feature table annotation. The central table **<annotationUniProtKB>** has been designed to store the following attributes of a PTM annotation: UniProtKB accession, start position, stop position, amino acid specific PTM type, evidence qualifier and source database. The business key of **<annotationUniProtKB>** may appear slightly strange in that all fields of this table, except for the surrogate primary key, are included; this business key allows for the duplication of a PTM annotation where all fields are equal except for the **<DbId>** and/or **<MethodId>**. This is used by the PTMDB import software to fully replicate a set of PTM annotations that are in the original import source, regardless of what is

already in the database. This allows for import sources to be compared with simple SQL statements. It should be noted that there are other obvious mechanisms that could be used to remove the requirement to duplicate data; such as creating a one-to-many relationship between **<annotationUniProtKB>** and additional mapping tables, i.e. **<annotationToDatabase>**.

The tables **<variantAnnotations>** and **<isoformAnnotations>** enable the PTMDB to store PTMs that do not apply to the protein primary sequence found in UniProtKB entries. The tables **<variants>** and **<isoforms>** are designed to store identifiers that can be looked up in the corresponding UniProtKB entry feature table, which will contain descriptions for how to transform the protein primary sequence into the variant or splice variant for the corresponding identifier.

The tables **<enzyme>** and **<enzymeAnnotation>** have also been created so PTM annotations can be associated with enzymes that are known to catalyse the corresponding PTM type at the corresponding residue.

The Perl class `<org.drd20.bioinformatics.database.ptmdb.PtmDbErd>` can be used to add PTMs to the PTMDB.

Section 2.5.2(a) Sequence space

To maintain consistency between different PTM annotations a single primary protein sequence is considered authoritative in the PTMDB. An imported PTM annotation must therefore be overlaid onto a protein sequence in the PTMDB. As most annotation databases can be mapped into UniProtKB this has been used to source such authoritative sequences. Upon update of the PTMDB, the latest version of UniProtKB is incorporated in the PTMDB as a series of tables in the namespace **<UniProtKB>**.

In UniProtKB there are two important attributes of a sequence, both of which the PTMDB supports, allowing for the verification of protein sequences.

1. Sequence version number – every time a sequence is changed in a UniProtKB entry by a UniProtKB curator, this number is incremented by one (Bairoch 2009).
2. Given protein primary sequence.

When importing an annotation into the PTMDB, at the very least, one of these two attributes should match. Note that, in addition to entries having a different sequence or version number, some may be completely missing. Missing entries may occur through the deletion, addition or merging of UniProtKB entries; exactly how this manifests itself depends on whether the PTMDB or the external database are using a newer version of UniProtKB. Figure 15 shows the simple routine that the PTMDB uses to validate protein sequences from external databases.

Section 2.5.2(b) *Taxonomy support*

Taxonomic trees have been included in the PTMDB to aid in the analysis of PTMs at different taxonomic levels. The NCB Entrez Taxonomy (NET) project distributes taxonomic trees for use by the scientific community (Sayers *et al.* 2009; Benson *et al.* 2009). These trees include the names of all organisms that have a protein or nucleotide in the genetic databases. Each UniProtKB entry is associated with a taxonomic identifier from the NET (Bairoch 2009). The NET taxonomic tree has been incorporated into the PTMDB using the tables shown in Figure 13.

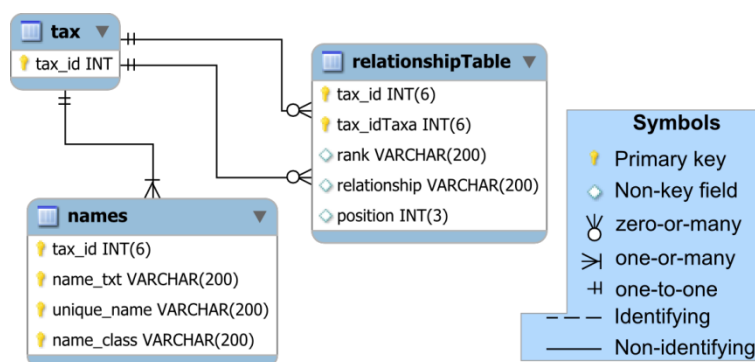


Figure 13: NCBi Entrez taxonomy tables in the PTMDB. These tables are stored in the *<ncbiEntrezTaxonomy>* namespace in the PTMDB. Note that the table *tax* isn't actually present in the PTMDB; it's included in this diagram so that relationship cardinality can be shown for the other two tables. These tables haven't yet been migrated from the MyISAM portion of the PTMDB schema to the InnoDB; therefore these foreign key constraints don't actually exist in the database.

The table *<names>* associates NET identifiers with all their synonyms; these synonyms are categorised by a class, stored in the field *<name class>*. The scientific name associated with a given NET identifier is categorised with the class "Scientific name". The table *<relationships>* stores the flattened taxonomic tree which can be used to identify all NET identifiers directly or

indirectly descended from another; for example to identify all mammalian species a search could be performed on this table by settings the field **<tax_idTaxa>** to the NET identifier corresponding to the NET node Mammalia.

This taxonomic tree can be imported into the PTMDB using the scripts `<ptmDB/ncbiEntrezTaxonomy/namesToMySQL.pl>` and `<ptmDB/ncbiEntrezTaxonomy/createRelationshipTable.pl>`. The latest version of this tree was imported into the PTMDB using the above scripts. This version contained 213,942 species in the super-kingdom Eukaryota, 103,800 for Bacteria and 3,432 for Archaea. In the taxonomic trees there were 620,000 taxonomic nodes.

No PTM annotation may be included in the PTMDB unless it has an associated taxonomic identifier that can be found in the NET.

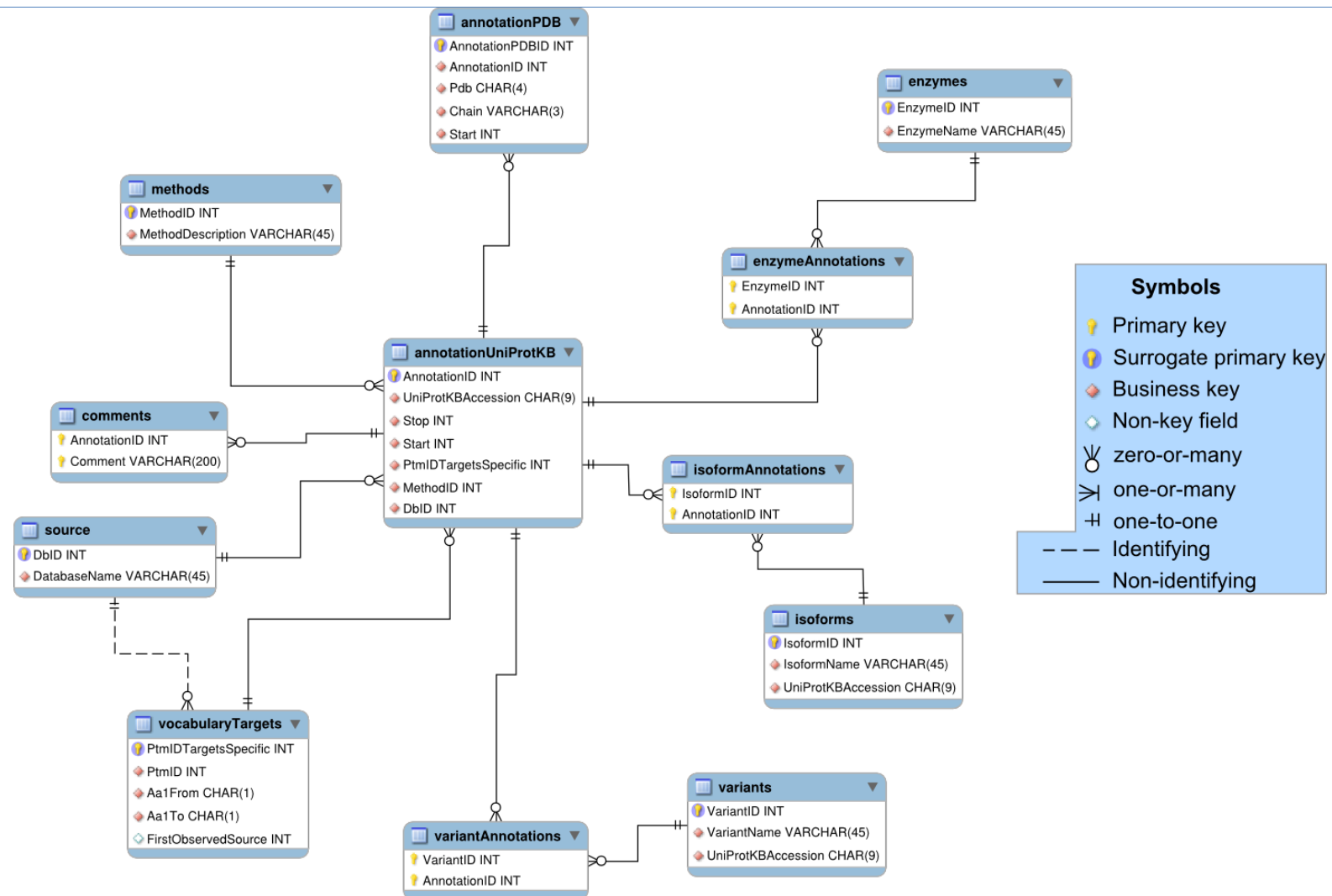


Figure 14: PTMDB annotation storage schema.

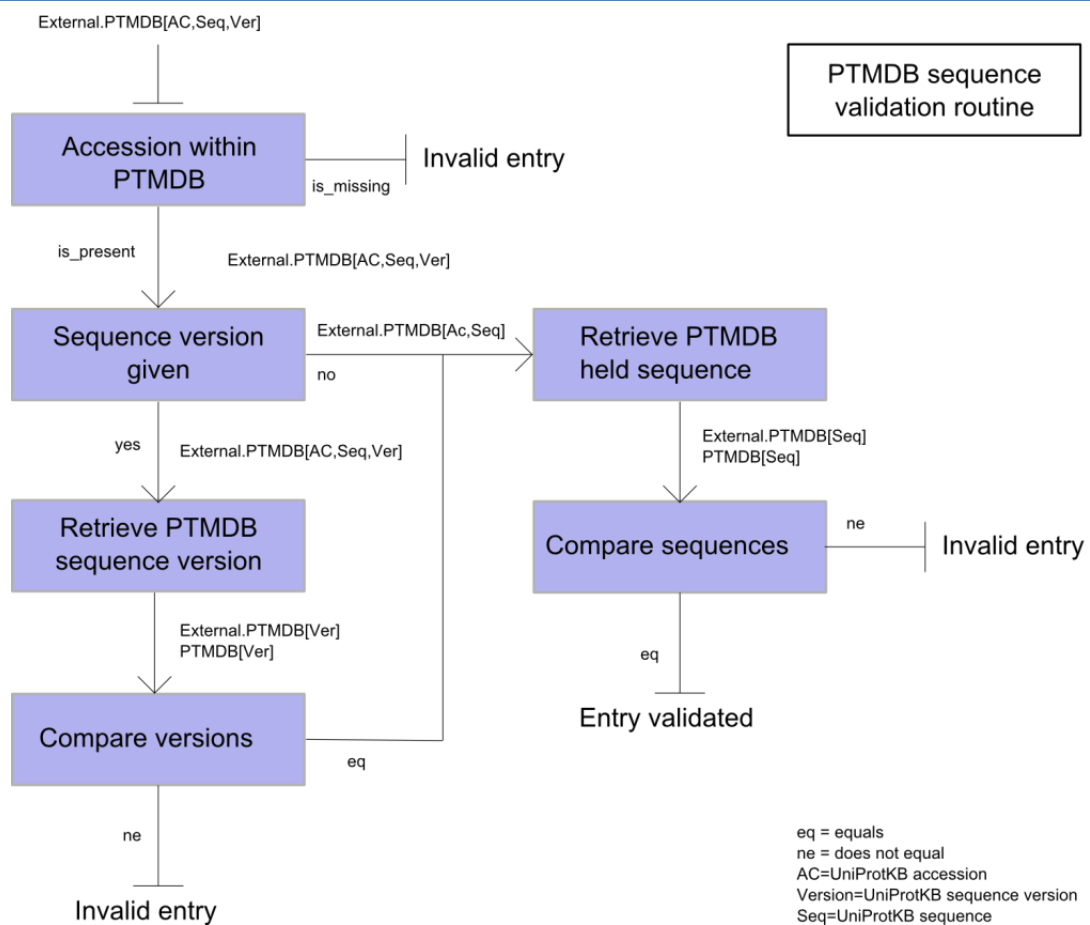


Figure 15: PTMDB protein primary sequence validation routine.

Section 2.5.3 UniProtKB import into the PTMDB

Software has been written that is able to import PTM annotations from the Swiss-Prot database into the PTMDB, using the schema and software library previously described. Additional information, which is not directly related to PTM annotations, can also be imported by this software to support other work presented in this thesis. Table 8 contains a list of the additional information that is imported into the PTMDB from the UniProtKB.

Swiss-Prot field ID	Description
AC	UniProtKB Accession
ID	UniProtKB Identifier
SQ	Protein Sequence
DT	Sequence revision number
OS	Taxon Name
GN	Gene Name
OG	Organelle
KW	Keywords
DR	Cross references to other databases

Table 8: Non-PTM annotation related UniProtKB fields imported into the PTMDB.

A Perl script has been designed that is able to parse UniProtKB flat files, extracting and uploading both the PTM annotations and generic UniProtKB fields shown in Table 8. This script makes use of the SwissKnife Perl API (Fleischmann *et al.* 2007) to parse UniProtKB entries and the Perl library <org.drd20.bioinformatics.database.ptmdb.PtmDbErd> to upload PTM annotations into the PTMDB. The upload script can be found here: <ptmDB/Main/UniprotUpload/uniprotToDatabaseErd.pl>. Note that UniProtKB is available for download as two flat files - one for Swiss-Prot and another for TrEMBL from the following URL <<http://www.uniprot.org/downloads>>.

Section 2.5.3(a) General statistics

Version 13.3 of the UniProtKB has been imported into the PTMDB to support the work presented in this thesis, this version consists of versions 55.3 and 38.3 of Swiss-Prot and TrEMBL respectively.

Swiss-Prot version 55.3 contains 366,226 entries. Table 9 shows that 15% of these entries have at least one PTM annotation. 282,200 residues are annotated with PTMs. The five most abundant PTM classes, ranked by number of proteins, are: Phosphorylation, N-linked Glycosylation, Lipidation, Acetylation

and Palmitation. Phosphorylation and Glycosylation account for 80% of the PTMs in Swiss-Prot, with 36.9% and 43.31%, respectively.

The distribution of PTM annotations by evidence qualifier, for the most abundant modifications is shown in Table 10 and Table 11. Table 10 shows the distribution based on the number of modified residues and Table 11 by the number of modified proteins. The values in Table 10 and Table 11 appear very similar; this occurs because most proteins only have one modification annotation for a given PTM class. The same results are shown graphically in Figure 16 and Figure 17. The majority of annotations either have the potential (predicted) or By similarity (cross-annotated) evidence types. The two most abundant PTMs, Phosphorylation and Glycosylation, show remarkably different evidence qualifier profiles. 41% of phosphorylation sites are backed up by experimental evidence; with the vast majority of remaining sites 56%, having been cross-annotated. In contrast, almost all glycosylation sites (94%) have been predicted using computational tools. The five PTM classes having the highest percentage of sites with experimental evidence are: Amidation (65%), Hydroxylation (57%), Pyrrolidone carboxylic acid (50%), Phosphorylation (41%) and O-linked glycosylation (33%). Note however that the number of actual annotations for some of these is quite low. For example there are only 995 O-linked glycosylation sites in contrast to 27,042 phosphorylation sites. It must also be pointed out that the Swiss-Prot curators do not transfer predicted annotations between protein entries (Farriol-Mathis *et al.* 2004); if they did the fraction of N-linked glycosylation annotations would most likely increase.

PTM Class	Proteins			Positions	
	Observed	Percentage		Observed	Percentage
		PTMDB	SP		
Swiss-Prot	56074		15.31	186408	
Phosphoprotein	23679	42.23	6.47	70360	36.99
Glycosylation	21844	38.96	5.96	85483	44.94
Glycosylation_N_Linked	21354	38.08	5.83	82381	43.31
Lipoprotein	5884	10.49	1.61	9132	4.80
Acetylation	5385	9.6	1.47	8319	4.37
Palmitate	3240	5.78	0.88	4411	2.32
Other	2409	4.3	0.66	2547	1.34
Methylation	2140	3.82	0.58	4559	2.40
Amidation	1966	3.51	0.54	2659	1.40
Pyrrolidone carboxylic acid	1120	2	0.31	1234	0.65
Myristate	1097	1.96	0.3	1099	0.58
Prenylation	931	1.66	0.25	1248	0.66
Glycosylation_O_Linked	866	1.54	0.24	2990	1.57
Pyruvate	759	1.35	0.21	759	0.40
Glycoprotein	757	1.35	0.21	757	0.40
GPI-anchor	756	1.35	0.21	756	0.40
Hydroxylation	390	0.7	0.11	2184	1.15
Sulfation	336	0.6	0.09	860	0.45
Flavoprotein	205	0.37	0.06	207	0.11
Peptidoglycan-anchor	197	0.35	0.05	197	0.10
FAD	139	0.25	0.04	139	0.07
Formylation	135	0.24	0.04	139	0.07
Covalent protein-RNA linkage	134	0.24	0.04	142	0.07
Gamma-carboxyglutamic acid	128	0.23	0.03	736	0.39
ADP-ribosylation	105	0.19	0.03	205	0.11
Nucleotide-binding	95	0.17	0.03	95	0.05
Hypusine	87	0.16	0.02	87	0.05
Citrullination	68	0.12	0.02	125	0.07
FMN	66	0.12	0.02	68	0.04
D-amino acid	58	0.1	0.02	81	0.04
Nitration	49	0.09	0.01	57	0.03
Organic radical	46	0.08	0.01	46	0.02
TPQ	42	0.07	0.01	43	0.02
Glycosylation_C_Linked	32	0.06	0.01	104	0.05
Oxidation	32	0.06	0.01	37	0.02
Bromination	21	0.04	0.01	24	0.01
S-nitrosylation	17	0.03	0	20	0.01
Glutathionylation	17	0.03	0	21	0.01
Covalent protein-DNA linkage	15	0.03	0	15	0.01
Glycosylation_S_Linked	8	0.01	0	8	0.00
Iodination	4	0.01	0	17	0.01
Selenium	4	0.01	0	4	0.00

Table 9: Distributions of PTM annotations imported from Swiss-Prot grouped by PTM class. Observed proteins: number of proteins in Swiss-Prot with at least one PTM annotation of the corresponding PTM class. Percentage protein: percentage of proteins in Swiss-Prot with at least one PTM annotation of the corresponding PTM class. Observed positions: Number of residues that are modified in Swiss-Prot database with the corresponding PTM class. Percentage positions: Percentage of modified sites which are of the corresponding PTM class. Note that a protein may receive Glycosylation PTMs of different linkage types; therefore the totals for the PTM class Glycosylation will not necessarily be the sum of the totals for the PTM classes, Glycosylation_[NCOS]_Linked.

PTM Class	Experimental	Probable	By similarity	Potential
Amidation	65.1	3.65	22.41	8.84
Hydroxylation	57.03	4.21	36.88	1.88
Pyrrolidone carboxylic acid	50.16	4.62	42.54	2.67
Phosphoprotein	41.19	0.73	56.5	1.58
Glycosylation_O_Linked	33.28	6.89	30.2	29.63
Acetylation	26.69	1.68	71.33	0.30
Sulfation	22.09	3.95	30.7	43.26
Methylation	19.34	2.15	76.34	2.18
Other	11.63	0.55	87.31	0.51
Myristate	10.92	2.18	59.05	27.84
Prenylation	9.08	1.53	83.28	6.11
Lipoprotein	6.81	6.26	53.59	33.34
Palmitate	5.33	8.91	50.12	35.64
Glycosylation	5.08	0.74	2.2	91.97
Glycosylation_N_Linked	4	0.5	1.14	94.36
Glycoprotein	3.83	3.04	15.72	77.41
Total	16.09	1.18	23.27	59.46

Table 10: Percentage of PTM sites imported from Swiss-Prot grouped by PTM class and evidence qualifier.

PTM Class	Experimental	Probable	By similarity	Potential
Amidation	66.22	4.55	24.09	5.15
Hydroxylation	54.05	2.7	38.82	4.42
Pyrrolidone carboxylic acid	52.22	4.63	41.19	1.96
Phosphoprotein	37.96	1	59.14	1.90
Sulfation	30.24	3.41	38.54	27.80
Acetylation	29.68	2.38	67.49	0.45
Glycosylation_O_Linked	25.48	7.82	46.14	20.56
Methylation	16.54	2.66	76.96	3.85
Myristate	10.76	2.19	59.16	27.89
Prenylation	9.88	2.04	80.67	7.41
Other	9.83	0.41	89.22	0.54
Glycosylation	7.85	1.57	4.22	86.37
Glycosylation_N_Linked	7.41	1.27	2.47	88.85
Lipoprotein	6.65	7.02	45.78	40.56
Palmitate	4.64	10.38	38.73	46.25
Glycoprotein	3.83	3.04	15.72	77.41
Total	19.02	2.21	33	45.77

Table 11: Percentage of PTM proteins imported from Swiss-Prot grouped by PTM class and evidence qualifier.

Number of distinct PTM sites grouped by class and experimental evidence

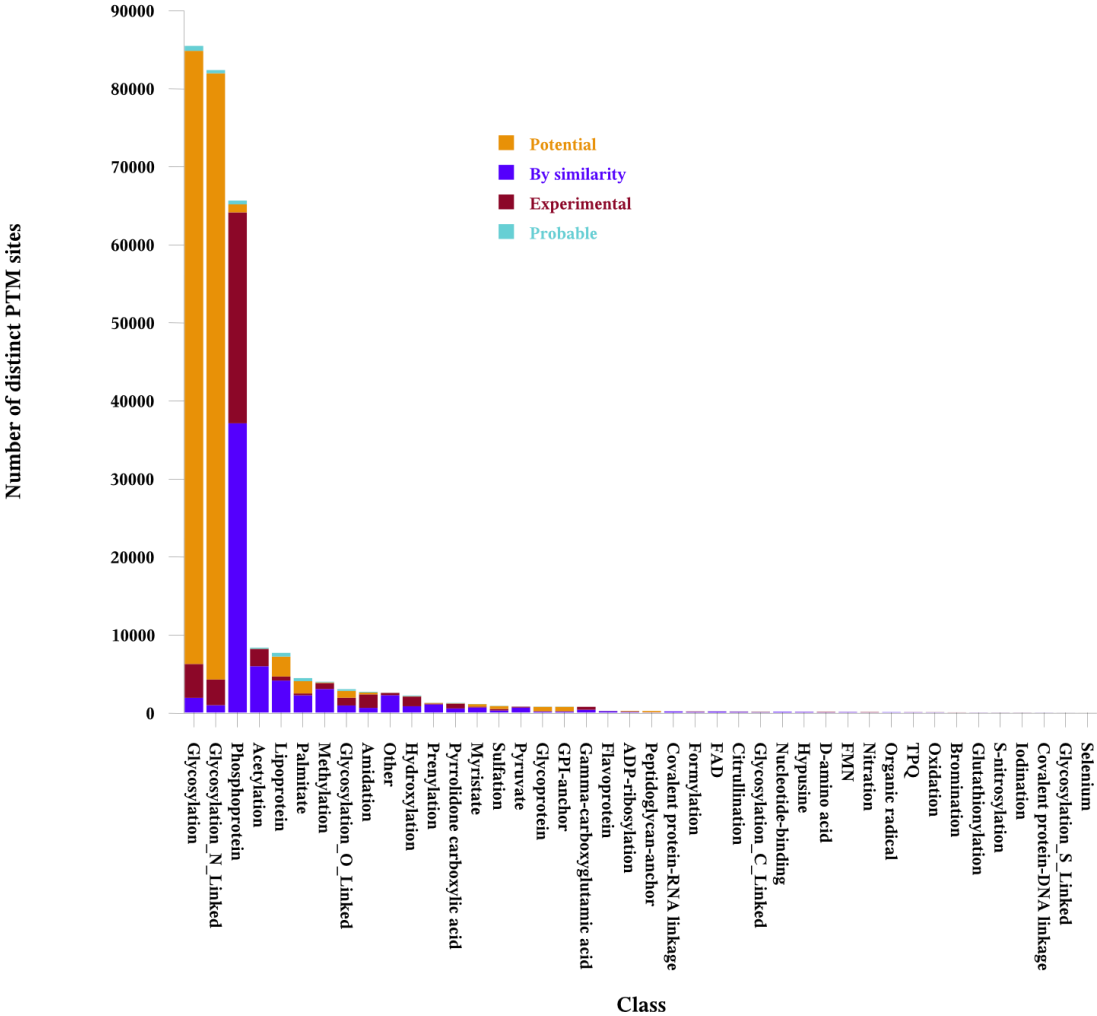


Figure 16: Number of modified residues grouped by PTM class and evidence qualifier

Section 2.5.4 Phospho.ELM

Phospho.ELM is a database of experimentally determined phosphorylation sites, detected by both HTPs and LTPs (High/Low Throughput Protocols). This database is maintained as part of the ELM (Eukaryotic Linear Motif) (Puntervoll *et al.* 2003) project; ELM contains a collection of small protein functional motifs. Although the PTMDB imports a vast number of phosphorylation sites from UniProtKB it should not be considered the authoritative resource for PTM data. Indeed UniProtKB, as previously discussed, only annotates PTMs to proteins that are in Swiss-Prot. In contrast, Phospho.ELM contains annotations on proteins in both Swiss-Prot and TrEMBL.

Phospho.ELM is available for download as a flat file from the following URL: <http://phospho.elm.eu.org/dataset.html>. The Perl class `<org.drd20.bioinformatics.database.ptmdb.Phospho.ELM>` has been created that is able to parse Phospho.ELM flat files and upload the extracted PTM annotations into the PTMDB using the previously described Perl class `<org.drd20.bioinformatics.database.ptmdb.PtmDbErd>`. This software is able to extract the following information from a Phospho.ELM flat file: UniProtKB accession, sequence version number, PTM type (Phospho.ELM uses the Swiss-Prot PTM vocabulary), residue number, evidence qualifier and enzymes that are responsible.

Not all Phospho.ELM annotations include the UniProtKB sequence version number; however all include the protein primary sequence. The sequence validation procedure shown in Figure 15 is used to verify that Phospho.ELM annotations have been made on identical sequences to those stored in the PTMDB (from UniProtKB version 13.3).

Phospho.ELM uses two evidence qualifiers, HTP and LTP, all PTM annotations in Phospho.ELM would qualify for the Swiss-Prot evidence qualifier Experimental. However it was decided that it might at some point become useful to be able to differentiate between those annotations determined by high versus low throughput techniques; therefore two new evidence qualifiers have been added to the PTMDB – Phospho.ELM_HTP and Phospho.ELM_LTP.

Section 2.5.4(a) *General statistics*

Version 7 of the Phospho.ELM database was parsed and uploaded into the PTMDB. It's important to note how many annotations were rejected by the PTMDB upload software because of sequence differences or UniProtKB entries simply being missing from the PTMDB; summary statistics are shown in Table 12. 879 Phospho.ELM instances were rejected by this procedure: 422 sequence mismatches, 426 missing UniProtKB accessions in the PTMDB and 31 UniProtKB sequence version mismatches.

	No Phospho.ELM instances
Sequences do not match	422
UniProtKB accession missing	426
Sequence versions do not match	31
Total	879

Table 12: Breakdown of the reasons for Phospho.ELM annotation rejection. Not all Phospho.ELM annotations contain the UniProtKB sequence version number; where these are missing the listed protein primary sequence (in the Phospho.ELM flat file) are compared with that in the PTMDB (taken from UniProtKB version 13.3). There are two possible reasons that UniProtKB accessions annotated in Phospho.ELM may be missing from the PTMDB; a) The PTMDB uses a newer version of UniProtKB than Phospho.ELM in which the missing accessions have been removed, or b) Phospho.ELM uses a newer version of UniProtKB than the PTMDB, which contains new accessions not found in the older version.

A detailed breakdown of the phosphorylation sites which were imported into the PTMDB from Phospho.ELM are shown in Table 13. Separate statistics are displayed in Table 13 (a) and Table 13 (b) for proteins from Swiss-Prot and TrEMBL respectively. 15,590 instances of phosphorylation sites were imported, 10,902 of which match instances in the UniProtKB. 2,036 new phosphoproteins have also been imported into the PTMDB.

Section 2.5.5 Negative PTM annotations

In science a negative conclusion or result should be given equal weight to a positive one. It is therefore important that bioinformatics databases record, and tools report, such negative experimental results. A negative PTM annotation results from experimental evidence that a predicted PTM cannot be detected (Farriol-Mathis *et al.* 2004). Farriol-Mathis *et al.*, (2004) emphasize that negative PTM events should be plausible; the example quoted is that the protein to be modified should be present within the same cellular compartment as the necessary PTM enzyme(s). To date the only database that maintains a list of negative PTM annotations is Swiss-Prot; they were originally intended to

be used as a negative dataset to train/benchmark PTM prediction tools (Farriol-Mathis *et al.* 2004).

Swiss-Prot records negative PTM annotations in the feature table of an entry using the feature key SITE; this key is used for any single amino acid feature which doesn't have a dedicated key of its own (Bairoch 2009). Negative PTM annotation lines can be detected by the presence of the word Not in field five of a feature table SITE line. The Perl class <org.drd20.bioinformatics.database.NotPtm> has been created that can extract such negative PTM annotations from Swiss-Prot flat files.

Phospho.ELM SwissProt results				
Type	New sites	New proteins	Matching sites	Matching proteins
Phosphoserine	2606	351	8196	3623
Phosphothreonine	656	147	1518	1284
Phosphotyrosine	777	124	1180	875
Phosphohistidine	1	1	0	0
Total	4040	510	10894	4407

(a)

Phospho.ELM TrEMBL results		
Type	New sites	New proteins
Phosphoserine	533	259
Phosphothreonine	82	68
Phosphotyrosine	41	28
Total	656	292

(b)

Table 13: Phospho.ELM import statistics. These tables show the number of phosphorylation sites and phosphorylated proteins imported into the PTMDB from version 7 of the Phospho.ELM database. (a) shows import statistics for those proteins which are in the Swiss-Prot section of the UniProtKB; Matching sites and Matching proteins shows the intersect between the Swiss-Prot and Phospho.ELM datasets in the PTMDB. (b) shows import statistics for proteins in the TrEMBL database (obviously there's no intersect counts).

Section 2.5.5(a) General statistics

Negative PTM annotations from version 55.3 of the Swiss-Prot database have been uploaded into the PTMDB; using the previously mentioned class. During the development of this class it was noted that not all of the PTM types listed for negative annotations conform to the previously described Swiss-Prot PTM vocabulary. Table 14 contains a list of these additional PTM types, along with the corresponding PTM type that has been registered with the PTMDB vocabulary. All of these annotations have been given the value Swiss-Prot-Not for the database name attribute in the PTMDB; so that these annotations can

easily be differentiated from all others. Some of these negative annotations were specific to particular variants; these have been given the value Swiss-Prot-Not-Var for the database name attribute.

PTM Listed	PTM registered in the SP-PTM-CV
Glycosylation	Glycosylation_Generic
Hydroxylation	Hydroxylation_Generic
N6-methylated	Methylation_Generic
Methylated	Methylation_Generic
Acetylated	Acetylation_Generic
Phosphorylated	Phosphoprotein_Generic
N-palmitoylation	Palmitate_Generic
S-palmitoylated	Palmitate_Generic
N-formylated	Formylation_Generic
Sulphated	Sulfation_Generic

Table 14: Additional PTM types created for the Swiss-Prot negative annotations.

Figure 18 displays the distribution of negative PTM annotations in the Swiss-Prot database grouped by PTM class. The current dataset is extremely small consisting of 135 experimentally verified negative PTM events. The largest negative datasets are those for Glycosylation (54) and Methylation (24).

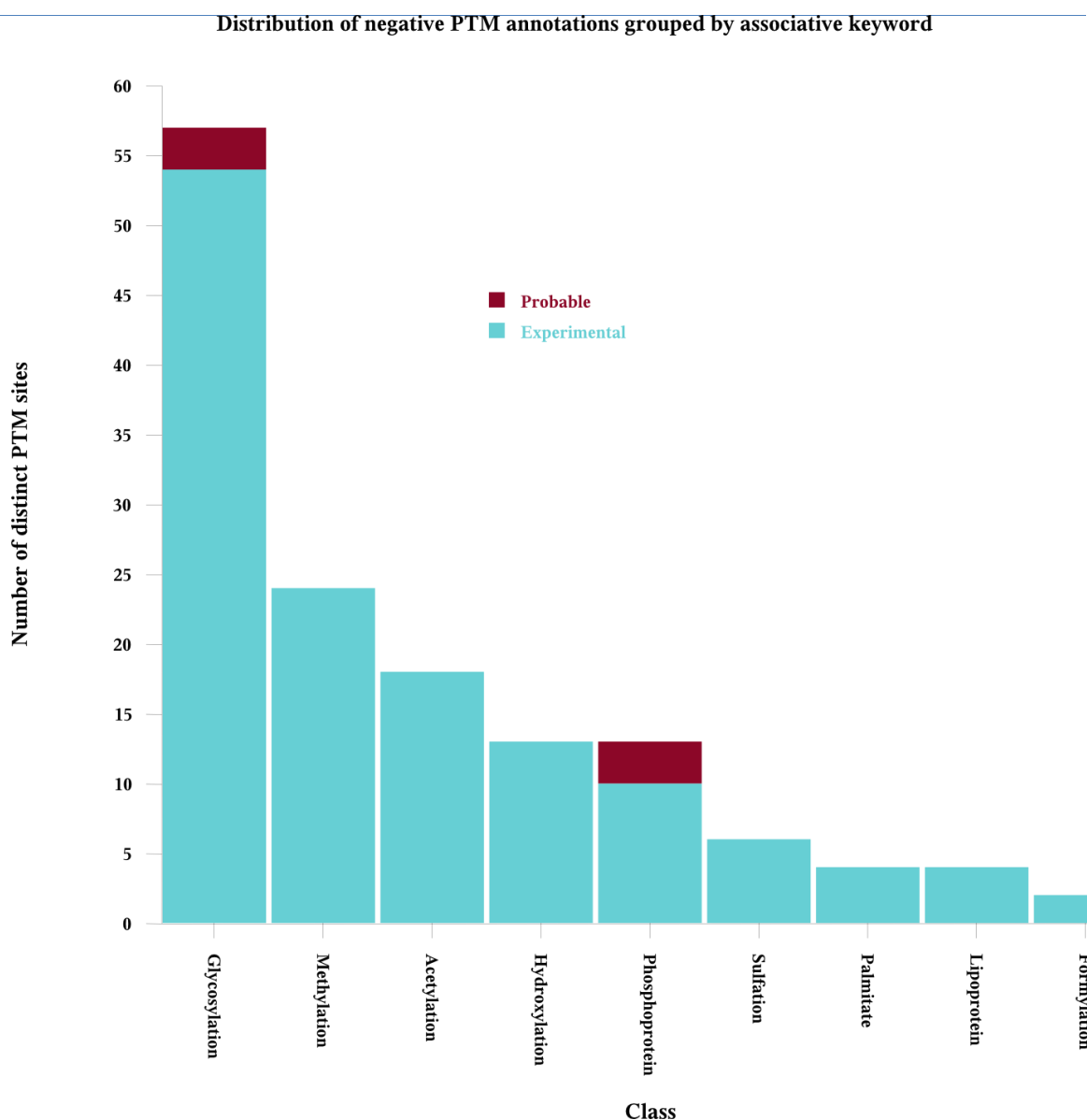


Figure 18: Distribution of negative PTM sites by PTM class.

Section 2.6 Glycan structure import using the PDB2LINUCS tool

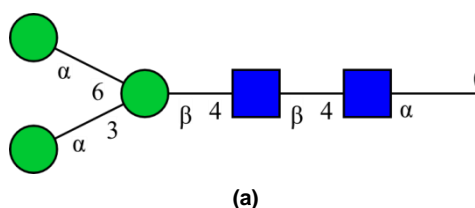
UniProtKB provides the following information regarding Glycosylation: modified residue, linkage type, terminal reducing sugar residue (where known) and mark-up to indicate if the glycan has additional sugars attached (see Figure 8 for further details). Therefore in the PTMDB none of the glycosylation site annotations have associated glycan structures. It was considered possible that glycan structures might vary between orthologues; thus being important to the study presented in this thesis.

There are currently only two databases that contain both glycan primary sequence and associated attachment points: GlycoSuiteDB and the PDB (Protein Data Bank). At the time this work was carried out, GlycoSuiteDB was a commercial product, thus precluding the inclusion of data from this database in the PTMDB. The PDB contains some protein structures with covalently bound glycans; as such both the identity and connectivity of sugar residues can be extracted from such PDB entries. An overview of two bioinformatic resources, GDB::Structures (Nakahara *et al.* 2008) and PDB2LINUCS (Lutteke, Frank, and von der Lieth 2004), which can extract glycan structures from the PDB, is shown in Table 15.

	GDB::Structures	PDB2LINUCS
Verifies glycan structures	Yes	Yes
Corrects glycan structures	No	Yes
Database of glycan structures	Yes	Yes
Accepts PDB formatted files	Yes	Yes
Displays glycan structure	IUPAC	LINUCS
Displays attachment position	No	Yes
Links to other databases	KEGG	CFG

Table 15: Comparison of tools that can extract glycan structures from the PDB.

GDB::Structures doesn't indicate which residue the extracted glycan structures are attached to; therefore PDB2LINUCS has been chosen to incorporate glycan structures into the PTMDB. Note that this resource is so named because it outputs glycan structures in the LINUCS (Linear notation for unique description of carbohydrate structures) format (Bohne-Lang *et al.* 2001); see Figure 19 for an example glycan structure in LINUCS format.



`[] [ASN] {[(4+1)] [a-D-GlcpNAc] {[(4+1)] [b-D-GlcpNAc] {[(4+1)] [b-D-Manp] {[(3+1)] [a-D-Manp] {} [(6+1)] [a-D-Manp] {} } } } }`

(b)

Figure 19: Core N-linked glycan in CFG symbolic nomenclature and LINUCS format. (a) The GlycoWorkbench tool has been used to produce a graphical representation of the LINUCS code shown in (b).

Section 2.6.1 Annotation extraction

The PDB2LINUCS web site (<<http://www.glycosciences.de/tools/pdb2linucs/>>) allows users to browse glycan structures and their attachment sites using a mirror of the PDB created in 2003. Two issues are caused by the PDB2LINUCS web site using a relatively old version of the PDB: a) new structures deposited since 2003 can't be found with this tool, and b) it's time consuming to map a residue in a PDB structure from an old version of the PDB into a much newer version of UniProtKB (for reasons that will become clear shortly). Fortunately the PDB2LINUCS web site allows for glycan structures to be searched for in a PDB file that the user uploads. Note that the PDB2LINUCS resource doesn't explicitly provide a web service for Bioinformatic tools, however one of the original authors of the tool made it clear that they didn't mind their tool being used automatically by scripts.

Unless stated otherwise all of the following scripts and classes required to import PDB2LINUCS Glycosylation annotations into the PTMDB can be found in the directory/namespace <org.drd20.bioinformatics.database.ptmdb.GlycoSciences>. A number of scripts and classes will be described that carry out the process of importing and classifying these Glycosylation annotations; note that the single wrapper script <importPDB2LINUCS.pl> has been created which can carry out the whole import process.

The Perl script <[batchRetrievalRemediatedFiles.pl](#)> has been designed to submit PDB files from a local archive to the PDB2LINUCS web page; extracting the LINUCS ID (numeric identifier for the LINUCS code), LINUCS code, PDB ID, chain ID, and attached residue number. The appropriate command to send to the PDB2LINUCS web site was reverse engineered by analysing the source code of the corresponding user input pages; for reference the URL used by this script is <<http://www.glycosciences.de/tools/pdb2linucs/pdb2linucs.php>> with the parameters: notation="linucs", usechime="off" and userfile="encoded PDB file".

When this script was originally designed the PDB2LINUCS tool only outputted the residue number for amino acid residues with attached glycans; as will be shown below the chain ID is required to map residues from the PDB to

UniProtKB. At the time the only available solution to obtain the chain ID was to make use of information stored in a log file that PDB2LINUCS automatically generates on PDB file upload. The location of the log file has to be extracted from the HTML that the tool initially returns to the script. This log file contains the residue number and chain ID corresponding to the terminal-reducing end sugar (i.e. the sugar residue attached to the amino acid residue). Each glycosylation annotation, which is extracted from the raw PDB2LINUCS output, therefore consists of the following fields: terminal-reducing end sugar residue number and associated chain ID, and the residue number for the amino acid residue the corresponding terminal-reducing end sugar is attached to. LINK records in a PDB entry indicate linkages between residues that aren't implied by the primary sequence; this includes linkages between terminal-reducing end sugars and the amino acid residues they are attached to. The triplet of fields previously described can be searched for in the LINK records of the corresponding PDB file to identify the missing chain ID. This issue has since been brought to the attention of one of the authors of the PDB2LINUCS tool; the latest version of PDB2LINUCS includes chain identifiers for amino acid residues with attached glycans (note that the above script has since been adjusted to support this new information).

Section 2.6.2 PTMDB import process

The script `<batchRetrievalRemediatedFiles.pl>` produces a list of residues (identified by PDB ID, Chain ID and residue number) and the glycan structures (in LINUCS format along with a LINUCS ID) that they are attached to. These residues must be mapped onto the protein primary sequences stored in the PTMDB for each corresponding UniProtKB entry. A database called PDBSWS has been published which contains a map between residues in the PDB and their equivalent residue numbers in the UniProtKB (Martin 2005). Note that a single PDB file may contain multiple chains; these chains may represent different protein/polypeptide chains. It's for this reason that it's necessary to know the chain ID in addition to the residue number (of glycosylated residues in a PDB file) to identify the correct UniProtKB entry. The PDBSWS database is available for download as a CSV file from the following URL: `<http://www.bioinf.org.uk/pdbsws/pdbsws_res.txt>`. A simple SQL script has

been created that incorporates this mapping directly into the PTMDB; <createMySQLMapping.sql>. The output from the script <batchRetrievalRemediatedFiles.pl> can be directly fed into the script <dumpResultsToDatabaseErd.pl>; this additional script maps the PDB glycosylation sites into the UniProtKB (using the PDBSWS) and uploads the new annotations into the PTMDB using the <PtmDbErd> class. The PTM type is set to the LINUCS ID returned by the PDB2LINUCS tool; in the vocabulary schema this means that the field <vocabulary.PtmDescription> is being used to store LINUCS IDs; see Section 2.6.4(c) for further details on how LINUCS structures are supported in the PTMDB vocabulary.

Section 2.6.3 Import into the PTMDB

A snapshot of the PDB was downloaded and archived on the 23rd May 2008; this version contained 50,830 structures compared to 22,448 in the version of the PDB used to generate the default PDB2LINUCS dataset (available on their web site). The script <batchRetrievalRemediatedFiles.pl> was used to upload all of the 50,830 PDB files to the PDB2LINUCS tool; note that this script includes sleep code, so not to overload the PDB2LINUCS web server. The script <dumpResultsToDatabaseErd.pl> was then used to upload the results into the PTMDB.

The original PDB2LINUCS dataset contained 5,647 glycan chains in 1,663 PDB entries. 1,457 of the 50,380 entries in the new dataset contained protein-glycan covalent linkages. The PDB2LINUCS tool returned 422 glycosylation sites without a LINUCS ID; these structures have been removed from the results returned by the tool. Manual inspection of the structures without a LINUCS ID suggested that the tool was unfamiliar with some of the sugars they contained. 6,371 unique quadruplets (PDB ID, chain ID, residue number, LINUCS ID) were returned by the PDB2LINUCS tool. Mapping these sites into the UniProtKB (with the PDBSWS database) revealed that 3,755 of them mapped to the same triplet formed of: UniProtKB accession, residue number, and LINUCS ID. In total 1,363 glycosylation sites (unique pairs formed of a UniProtKB accession and residue position) have had annotations imported for them from the PDB. These sites are distributed amongst 540 UniProtKB accessions. 271 different

LINUXS encoded structures were represented in the results returned by the PDB2LINUXS software.

Section 2.6.4 Glycan classification

A manual inspection of the glycan structures found by the PDB2LINUXS tool appeared to reveal that the vast majority were short truncated structures. In order to investigate this further it was decided that it would be useful to cluster glycan structures in some way. The first step in this process involved splitting the structures into those which were n-linked and those which were o-linked; this can be done using the script `<generateUniqueGlycanList.pl>`

Section 2.6.4(a) N-linked

N-linked glycan structures are usually classified as being of one of the following types: high mannose, complex, hybrid, or core (Walsh 2006). The glycan structure database which can be found at the following URL: http://www.glycosciences.de/sweetdb/start.php?action=form_class_nglycan includes class designations for n-linked glycans; these designations have been used to manually derive the simple rules used to cluster them. The first class core is formed of the core n-linked penta-saccharide (formed of two GlcNAc and three Man residues); see the glycan in Figure 20 labelled core3.

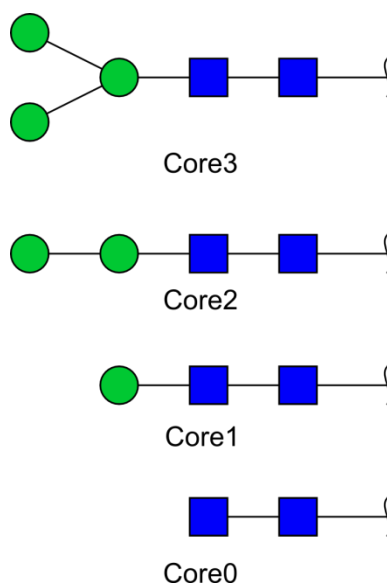


Figure 20: N-linked glycan core types used in the classification of structures from the PDB.

An example of each of the remaining three classes can be seen in Figure 2. A structure is classed as high mannose if the only residues present in the

branches initiated by the two antenna residues of the core penta-saccharide are mannose. A hybrid structure is one that has a mixture of mannose and non-mannose residues in its branches. Finally if all residues in the branches are non-mannose residue types the structure is considered to be complex. During the maturation process of an n-linked glycan the core penta-saccharide is trimmed back and elaborated upon with other sugars (Walsh 2006). A brief check of the structures imported into the PTMDB from the PDB revealed that many lacked the core penta-saccharide. To enable these structures to also be assigned a meaningful class the following classification system has been implemented. First the classification system identifies which one of the four core structures shown in Figure 20 are present; core3 is the core penta-saccharide, the remaining cores 2-0 have one successively less mannose residue. The system then uses the same rules previously outlined to identify which residues are present in the branch(s) that extend from the antenna terminal mannose residue(s); the traditional glycan class names are preceded by the core type found (e.g. n-linked-core3-high mannose).

This classification scheme can be carried out using the Perl class <GlycoSciencesNClassifier>.

Section 2.6.4(b) *O-linked*

As the number of o-linked glycans identified by the PDB2LINUCS tool was relatively low; no attempt has been made to classify them. For consistency all o-linked structures have been associated with the PTM class o-linked (specific to LINUCS o-linked PTM types). These PTM types have also been associated with the PTM classes Glycosylation and Glycosylation_O_Linked. The script <GlycanOLinkedTempClassifier.pl> attaches the o-linked PTM class to the o-linked structures output by the <generateUniqueGlycanList.pl> script.

Section 2.6.4(c) *Incorporation into the PTMDB vocabulary*

Both the o-linked and n-linked classifiers output their classification results in the same format. The script <insertLinucsIntoVocabulary.pl> is able to read these classification results and upload them into the PTMDB vocabulary (using the class <VocabErd>). Figure 21 contains those tables of the PTMDB vocabulary that are relevant to the following discussion on how this was achieved. First

note that the field **<vocabulary.PtmDescription>** has an upper size limit of 200 characters (this length was chosen based on the field and index size limits imposed by the **InnoDB** storage engine); this isn't enough characters to store many of the LINUCS patterns returned by the PDB2LINUCS tool. It was therefore decided that the glycan structures would be represented in the PTMDB vocabulary by their LINUCS IDs (stored in the **<vocabulary.PtmDescription>** field); these thus become new PTM types in the vocabulary. The LINUCS patterns are connected to their corresponding PTM types using the table **<vocabularyToGlycanStructure>**. The classes created by the n-linked and o-linked classifiers are added to the table **<keywords>** and associated with their corresponding PTM types using the table **<vocabularyToKeywords>**.

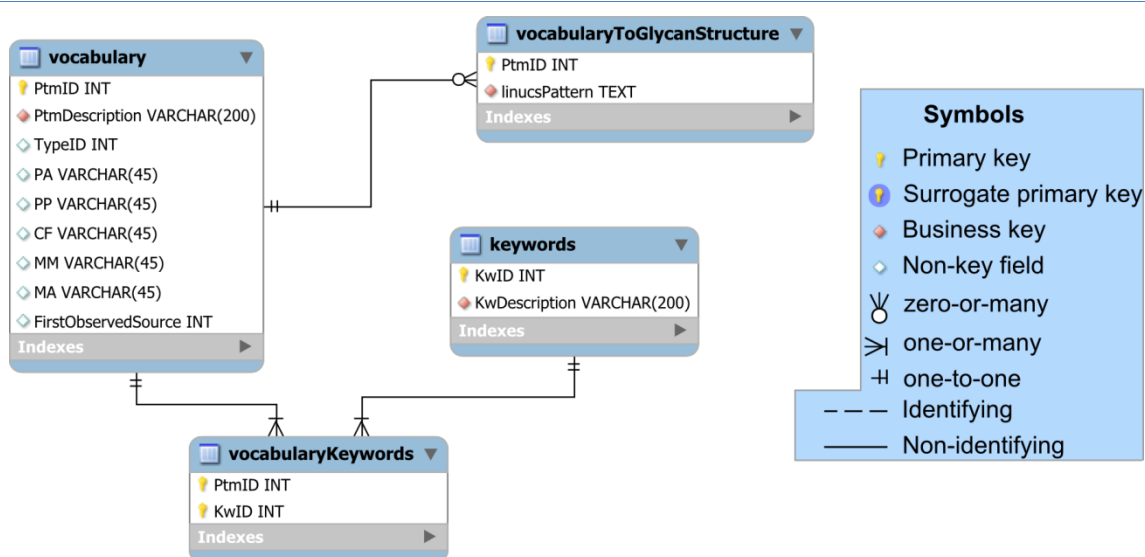


Figure 21: Glycan classification support in the PTMDB vocabulary

Section 2.6.4(d) Classification summary

Table 16 summaries the number of PDB and UniProtKB entries and individual glycosylation sites by the glycan classifications previously described. N-linked glycans were the most abundant accounting for 86% of those structures imported. Of the 235 n-linked glycan structures 41 were core0, 24 core1, 82 core2 and 88 core3. 123 (52%) of the structures did not have extensions to the core structure that was present. The largest single class of structures was core0 with 41 entries; note that other categories had a similar number of structures (e.g. core2 35, core3 High mannose 33). Of the extended classes

High mannose was associated with the most structures (61 structures). 509 UniProtKB entries were found to be represented in the Glycosylated PDB entry set. 466 of these had core0 structures associated with them (not necessarily exclusively though). 1,236 residues in the UniProtKB were found to correspond to the 5,483 residues glycosylated in the PDB dataset. 79% of the glycosylated residues were associated with core0 class structures (again not necessarily exclusively).

	PDB			UniProtKB	
	Structures	Entries	Sites	Entries	Sites
n-linked[core0]	41	1260	4480	466	1081
n-linked[core1]	22	238	463	116	174
n-linked[core1] Complex	2	19	22	3	4
n-linked[core2]	35	94	129	51	62
n-linked[core2] Complex	19	24	25	14	15
n-linked[core2] High mannose	28	69	82	26	27
n-linked[core3]	25	82	123	44	51
n-linked[core3] Complex	29	49	72	19	20
n-linked[core3] High mannose	33	69	86	32	37
n-linked[core3] Hybrid	1	1	1	1	1
o-linked	36	128	688	48	127
Glycosylation_N_Linked	235	1383	5483	509	1236
Glycosylation_O_Linked	36	128	688	48	127
Glycosylation	271	1457	6171	540	1363

Table 16: Glycan classification in the PDB2LINUCS dataset imported into the PTMDB. Note that the counts for *n-linked[core0]*, *n-linked[core1]*, *n-linked[core2]*, and *n-linked[core3]* are not inclusive of the extended classes with the same core type; i.e. the counts for *n-linked[core1]* do not include the counts for *n-linked[core1] complex*. In addition note that the counts for PDB entries, PDB sites, UniProtKB entries and UniProtKB sites for the class *Glycosylation* will not necessarily be equal to the totals for the values for the classes *n-linked**, and *o-linked*.

36 o-linked structures have been imported which mapped to 48 UniProtKB entries and 127 residues.

The PTMDB of course already contains a large number of Glycosylation sites imported from the Swiss-Prot database. Table 17 displays the intersect between the Swiss-Prot glycosylation annotations and those in the PDB2LINUCS dataset. Of the 509 UniProtKB entries in the PDB2LINUCS dataset which have n-linked Glycosylation annotations; 386 have corresponding entries in the Swiss-Prot database which are also annotated with n-linked glycans. Of the 1,236 n-linked residues in the PDB2LINUCS dataset 936 correspond to n-linked residues in the Swiss-Prot database

Type	New		Matching	
	Sites	Entries	Sites	Entries
n-linked[core0]	45	13	824	358
n-linked[core1]	5	3	136	87
n-linked[core1] Complex	0	0	3	2
n-linked[core2]	2	0	46	38
n-linked[core2] Complex	0	0	12	11
n-linked[core2] High mannose	0	0	22	21
n-linked[core3]	1	0	45	39
n-linked[core3] Complex	0	0	15	14
n-linked[core3] High mannose	2	1	28	26
n-linked[core3] Hybrid	0	0	1	1
o-linked	54	28	67	15
Glycosylation_N_Linked	51	14	936	386
Glycosylation	105	31	1003	395

Table 17: Intersect between the Swiss-Prot and PDB2LINUCS imported Glycosylation annotations. Note that the same rules regarding the inclusivity of the counts in this table as explained in the caption of Table 16 apply.

Protein chains in a PDB file are not limited to being mapped to Swiss-Prot entries; the PDBSWS database also maps chains to residues in the TrEMBL database. As the only source of Glycosylation annotations imported into the PTMDB (other than from the PDB) is the Swiss-Prot database; all sites imported using the PDB2LINUCS tool which map to residues in TrEMBL entries are new. Table 18 displays the number of TrEMBL entries and residues which have had Glycosylation annotations imported into the PTMDB from the PDB2LINUCS dataset. 255 residues in TrEMBL entries are annotated with Glycosylation sites in the PDB2LINUCS dataset imported into the PTMDB; these map to 114 TrEMBL entries.

Type	New	
	Sites	Entries
n-linked[core0]	212	95
n-linked[core1]	33	26
n-linked[core1] Complex	1	1
n-linked[core2]	14	13
n-linked[core2] Complex	3	3
n-linked[core2] High mannose	5	5
n-linked[core3]	5	5
n-linked[core3] Complex	5	5
n-linked[core3] High mannose	7	5
o-linked	6	5
Glycosylation_N_Linked	249	109
Glycosylation	255	114

Table 18: Breakdown of the Glycosylation annotations for TrEMBL entries in the PDB2LINUCS imported dataset. Note that the same rules regarding the inclusivity of the counts in this table as explained in the caption of Table 16 apply.

Chapter 3

Incorporation of homology assignments into the PTMDB

Section 3.1 Summary

One of the main aims of creating the PTMDB was to create a resource that could be used to allow users to ask questions about the conservation of modifications between homologous proteins. Publically available databases such as KEGG and InParanoid contain orthologue and paralogue assignments between large numbers of species. Orthologue annotations from both these projects were not suitable for inclusion in the PTMDB as the protein sequence space analysed did not match that of the PTMDB. The software used to build the InParanoid dataset was obtained from the authors, but failed to produce any orthologue assignments. A new implementation of the InParanoid algorithm has therefore been created to populate the PTMDB with orthologue assignments – called CoPaO (Clusters of Paralogues and Orthologues). CoPaO has been designed to distribute orthologue detection around a compute-cluster running the SunGrid engine software.

This chapter begins with a short overview of protein evolution and the correct use of the terms orthologue and paralogue. Following on from these definitions is a brief overview of the three classes of orthologue detection algorithm: tree-based, graph-based and a hybrid of the previous two. A comparison of two graph-based algorithms, InParanoid and COG, is then presented. The CoPaO algorithm is then described along with the process used to remove redundancy from the PTMDB sequence space. The orthologue assignments detected by the CoPaO program between *H. sapiens* and *Mus musculus* are then compared to those in the original InParanoid dataset. As an additional validation step the gene ontology annotations associated with Orthologues have been compared.

Finally a brief discussion is presented on the conservation of the *H. sapiens* proteome in a selection species.

Section 3.2 *Protein evolution*

John Maynard Smith originally defined protein evolution as a walk around sequence space that doesn't pass through non-functional intermediates (Smith 1970). He also postulated that an increase in the number of genes in an organism's genome may occur by gene duplication (Smith 1970). Such a gene duplication event results in two identical copies of a gene in a genome – a condition that is unsustainable without an appropriate selection pressure to maintain the copy (He and Zhang 2005). As a result, the two genes naturally diverge. Two theories have been put forward to explain the divergence: neofunctionalisation and subfunctionalisation. Neofunctionalisation maintains that one gene alone is required to meet the demands of the organism, whereas the other is free to acquire mutations, leading to novel functions in the resultant protein (He and Zhang 2005). Subfunctionalisation is where the combination of the two genes lessens the selective pressure on either one alone, allowing both to diverge to a point where both are now required for the original function (He and Zhang 2005).

Two nucleotide sequences that have been produced by a duplication event share the same ancestor (Fitch 1970). Two sequences that share a single common ancestral form are referred to as being homologous (Fitch 1970). Such sequences are referred to as having undergone divergent evolution (Fitch 1970). Sequences that don't share a single common ancestor, but are nevertheless similar, are referred to as having undergone convergent evolution – such sequences are said to be analogous (Fitch 1970).

Section 3.2.1 Defining Homology

Homologous relationships between genes can be subdivided based on when the gene duplication occurred relative to the LCA (Last Common Ancessor) (Fitch 2000). Genes that originate from a single ancestral gene in the LCA are referred to as orthologues (Koonin 2005). Those related by gene duplication rather than speciation are referred to as paralogues (Koonin 2005). Paralogues

can be further categorised as in-paralogues, when the duplication occurred after speciation, and out-paralogues, when it occurred before (Koonin 2005).

An example is shown in Figure 22 of a single gene that undergoes duplication and subsequent speciation. In the first species gene “X” has been duplicated to produce X_1 and X_2 – these are in-paralogues. A subsequent speciation event produces two species “A” and “B”, which have copies of each of these genes. The relationship between the genes A_{X1} , B_{X2} and A_{X2} is out-paralogous – as their ancestry can be traced back to the original gene duplication event in species “X”. The inverse is also true – the genes A_{X2} , B_{X1} and A_{X1} are also out-paralogous to each other. Finally the ancestry of A_{X1} and B_{X1} can be traced back to the single ancestor protein X_1 (likewise B_{X2} and A_{X2} to X_2). These are therefore orthologues.

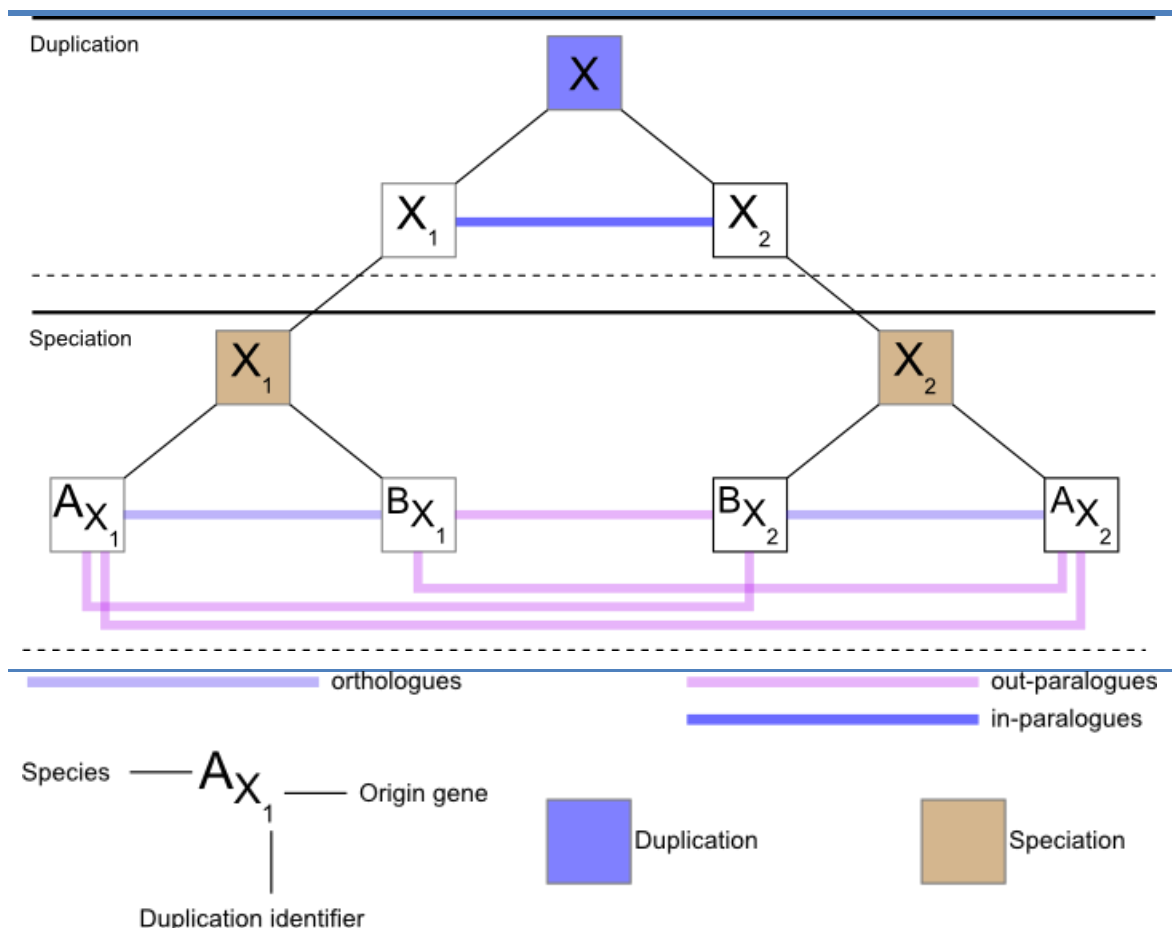


Figure 22: Simple model of a gene undergoing duplication and subsequent speciation. Gene X has been duplicated to produce X_1 and X_2 these two genes are in-paralogous to each other. A subsequent speciation event produces two species “A” and “B”, which both have copies of the X_1 and X_2 . A_{X1} and B_{X1} are orthologous to each other, as are B_{X2} and A_{X2} . A_{X1} , B_{X2} and A_{X2} are out-paralogues, as are A_{X2} , B_{X1} and A_{X1} .

Section 3.2.2 Horizontal Gene Transfer (HGT)

So far homologous genes have been referred to as being either orthologous or paralogous. It has been reported that a large number of bacterial genes appear to have been transferred between species (Daubin, Moran, and Ochman 2003). These genes can neither be referred to as orthologues nor paralogues. They are instead referred to as xenologues (Koonin 2005). It has recently been suggested that a bacterial species is defined as a collection of organisms that can frequently exchange genetic material by HGT (Daubin, Moran, and Ochman 2003).

Section 3.2.3 Mixed ancestry

A number of processes have been proposed that allow for gene duplication to occur, including unequal cross-over, non-homologous driven repair and retro transposition (Hurles 2004). Most of these processes have a degree to which they randomly pick start and end positions of a duplication. Therefore the likelihood of creating non-functional (at best) and lethal (at worst) new sequences is highly likely (Hurles 2004).

During evolution some genes may acquire new domains from other genes by duplication. Note that it is in this situation that one may refer to one gene as being a percentage homologous to another. Walter Fitch has been at pains to point out that homology only refers to a common ancestry and not to the degree of similarity between two sequences (Fitch 2000). Therefore if 40% of a gene resulted from the duplication of part of another gene – these two genes can be referred to as being 40% homologous (Fitch 2000).

Section 3.2.4 Detection algorithms

Orthologue detection techniques can be classified as being: tree-based, graph-based or hybrid (Kuzniar *et al.* 2008) – see Table 19 for some examples of each of these.

Tree based methods start by constructing a multiple sequence alignment of homologous sequences. A phylogenetic tree is then constructed from this multiple sequence alignment. This “gene tree” is then reconciled with the “species tree”, which may have a different topology. The differences between

these two trees can be accounted for by gene loss, gene gain (by HGT) and gene duplication. This method is susceptible to errors in any of the following: gene tree, species tree and multiple sequence alignment. There is not always a “species tree” available, and although methods have been created that do not require one, such methods are considered difficult to automate. (Kuzniar *et al.* 2008)

Graph-based methods start by defining “an operational definition of orthology” (Kuzniar *et al.* 2008). Most graph-based techniques assume that two orthologous sequences should be more similar to each other than to any other sequence in either the other genome or their own (Kuzniar *et al.* 2008). This is normally expressed in terms of detecting reciprocal BLAST best hit pairs (Kuzniar *et al.* 2008). Graph-based methods therefore detect orthologues by performing all-against-all BLASTs of their respective genomes (Kuzniar *et al.* 2008). Not all graph-based methods have the same ability to differentiate between orthologous and the two types of paralogous relationship (Kuzniar *et al.* 2008). Further methods usually differ in their ability correctly to label sequences in the presence of gene deletion and horizontal gene transfer (Kuzniar *et al.* 2008).

Method classification	Examples
Tree based	COCO-CL, HOPS
Graph based	COGs, InParanoid
Hybrid	Ensembl compara, PHOGs

Table 19: Example orthologue detection techniques classified as: tree-based, graph-based and hybrid (adapted from (Kuzniar *et al.* 2008)).

Graph-based methods are generally considered to be computationally less intensive than tree-based methods as they do not require the construction of phylogenetic trees (Remm, Storm, and Sonnhammer 2001). For this reason it was decided that the PTMDB would incorporate annotations from a graph-based method. A brief overview of two graph-based methods and one hybrid is now presented.

Section 3.2.4(a) COG (Tatusov, Koonin, and Lipman 1997)

The orthologue detection algorithm created by Tatusov, *et al.* for the COG (Clusters of Orthologous Groups of proteins) database is described in this section (Tatusov, Koonin, and Lipman 1997; Tatusov *et al.* 2000). Orthologues are detected using a graph-based technique, which clusters proteins together based on Blast Best Hits (BBHs). The authors of COG decided that it was best only to cluster proteins together based on BBHs formed between phylogenetically distant species. The first step in generating the COG database was to perform an all-against-all sequence comparison using BLAST. Paralogues are detected first as those proteins whose highest similarity score was obtained from a protein which was part of the same genome. Symmetrical and asymmetrical BBHs form edges between the compared proteins. Triangles of proteins are formed when three proteins from phylogenetically distant taxa are all connected by BBHs. Triangles that have a common edge are merged recursively

The first version of COG required the comparison of 17,967 prokaryote proteins. COG was last updated in 2003 when a Eukaryote companion database KOG was published (Tatusov *et al.* 2003).

Section 3.2.4(b) InParanoid (Remm, Storm, and Sonnhammer 2001)

The work of Remm *et al.* describes the construction of a new algorithm, which attempts to address some of the short-comings they had identified with the COG database (Remm, Storm, and Sonnhammer 2001). As previously stated, by definition a COG can only exist where orthologues are shared between three phylogenetically distant taxa. They argued that this precludes the representation of lineage specific functions in the COG database. InParanoid was also created because COG lacked support for Eukaryotes – although this was obviously addressed with the later release of KOG.

Dessimoz *et al.* also criticise COG for allowing asymmetrical BBHs to form orthologous relationships (Dessimoz *et al.* 2006). This has made COG susceptible to clustering out-paralogues in a COG (Dessimoz *et al.* 2006). Remm *et al.* state that a pair-wise comparison between two individual species

will provide different orthologous assignments to one carried out between many species all at once (Remm, Storm, and Sonnhammer 2001).

The InParanoid algorithm has been designed to allow users to enquire as to the conservation of proteins between two species. The central aim of this new algorithm was to prevent out-paralogues from being clustered into orthologous groups.

The InParanoid algorithm detects symmetrical BBHs from the results of all-against-all BLAST comparisons, implemented as four separate BLAST runs. First the proteome of species A is queried against the proteome of species B followed by the reverse. This process must be repeated in both orientations to correctly detect orthologues – consider the following example. The BBH of Protein A_1 from species A is protein B_2 of species B; however the true orthologue of A_1 has been lost from species B – which is evident as the BBH of protein B_2 is protein A_2 . Without running the comparison in both orientations the algorithm would incorrectly conclude that the orthologue of A_1 is B_2 ; if the only query performed was species A against the proteome of species B. Note that requiring symmetrical BBHs does not guarantee that correct orthologue assignments will be made, as the algorithm is still easily thrown off by the multitude of gene deletion events that can occur (both symmetrical and asymmetrical, with regards to the event occurring in both or one species). All-against-all BLAST comparisons are performed between the same proteomes (i.e. proteome A against A and proteome B against B) to identify in-paralogues. In-paralogues can be detected from the results of these comparisons as proteins that are more similar to a protein from the same species than to one in a different species. The fullset of comparisons that must be performed for the InParanoid algorithm are shown below.

Query	Hit
-------	-----

- | | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
| <ul style="list-style-type: none"> • Proteome A – Proteome A • Proteome A – Proteome B • Proteome B – Proteome A • Proteome B – Proteome B | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|

Briefly the InParanoid clustering routine works as follows – see the original paper by Remm, *et al.*, for further details. Insignificant BLAST hits are excluded using a bit score cut-off of ≥ 50 and an overlap threshold of $\geq 50\%$ of the longest sequence (hit or query). Symmetrical BBHs are identified and placed into descending bit score order (so that the most similar are processed first), these form the initial homologous seed clusters. In-paralogues are identified and added to each seed cluster in-turn; note that this process involves the merging, separating, and destruction of seed clusters with a lower bit score.

The InParanoid algorithm makes the assumption that orthologues are more distantly related to each other than they are to their respective in-paralogues. This is implemented as only allowing in-paralogues to be clustered if their bit-score to the iso-orthologue is \geq that between the orthologues. If an inter-genomic BBH pair contains a protein that has already been clustered as an in-paralogue, in another cluster, then rules are applied to decide what action to take (i.e. seed cluster deletion, merging, etc.).

InParanoid also generates a confidence value for each in-paralogue, which indicates the degree of similarity compared to that shown between the main orthologues. This confidence value is calculated using the equation shown in Figure 23. The most similar proteins in an InParanoid cluster are the main orthologues – this equation assigns these a score of 1. In-paralogues are assigned a score that is > 0 and < 1 .

Confidence equation:

$$\frac{\text{Bit}(\text{Species}[A|B]_{\text{InParalogue}}, \text{Species}[A|B]_{\text{orthologue}}) - \text{Bit}(\text{SpeciesA}_{\text{orthologue}}, \text{SpeciesB}_{\text{orthologue}})}{\text{Bit}(\text{Species}[A|B]_{\text{orthologue}}, \text{Species}[A|B]_{\text{orthologue}}) - \text{Bit}(\text{SpeciesA}_{\text{orthologue}}, \text{SpeciesB}_{\text{orthologue}})}$$

Main orthologue self against self	$\text{Bit}(\text{Species}[A B]_{\text{orthologue}}, \text{Species}[A B]_{\text{orthologue}})$
Main orthologue against in-paralogue.	$\text{Bit}(\text{Species}[A B]_{\text{InParalogue}}, \text{Species}[A B]_{\text{orthologue}})$
Distance between main Orthologues	$\text{Bit}(\text{SpeciesA}_{\text{orthologue}}, \text{SpeciesB}_{\text{orthologue}})$

Figure 23: InParanoid confidence value equation for in-paralogues. Main orthologues receive a score of 1 and in-paralogues receive a score that is > 0 and < 1 .

The InParanoid program is available on request and datasets are periodically made available on the InParanoid web site. The InParanoid algorithm has also been used to create the OrthoDisease database, which connects genes from

model organisms to their *H. sapiens* orthologues and overlays OMIM (Online Mendelian Inheritance in Man) annotations onto these orthologous pairs (O'Brien, Westerlund, and Sonnhammer 2004).

Section 3.2.4(c) PHOG (Merkeev, Novichkov, and Mironov 2006)

Hybrid methods use phylogenetic trees to guide graph based clustering techniques (Kuzniar *et al.* 2008). The PHOG (Phylogenetic Orthologous Groups) technique is an interesting hybrid method that produces orthologous groups at each node of the taxonomy tree (Merkeev, Novichkov, and Mironov 2006). Briefly this technique starts with the creation of orthologous clusters, using a graph-based approach, between species at the same taxonomic node. These clusters are merged into what is termed a supergenome. Supergenomes, which are at the same taxonomic node, then undergo the same orthologous clustering procedure. This process is repeated until the root of the taxonomic tree has been reached.

Section 3.2.5 Summary

None of the existing databases of orthologues are particularly well suited for inclusion in the PTMDB because most have either been abandoned or are updated very infrequently. Infrequent updates are a problem because it: a) excludes newly sequenced genomes and b) increases the probability of sequences being present in one database but missing in the other. For example the COG database hasn't received an update since 2003, and the InParanoid database was last updated in 2006. It was therefore decided that new orthologue assignments would need to be made specifically for the PTMDB. With limited computational resources it was decided that a graph-based technique should be used for the detection of orthologues in the PTMDB. The advantages of the InParanoid algorithm over that used to create COG have already been discussed. The InParanoid authors' implementation of their algorithm was obtained, but failed to produce any orthologue assignments. Rather than attempt to fix their software it was decided that a new implementation of the InParanoid algorithm would be created. This new

implementation has been designed to be able to take advantage of clusters running the SunGrid engine software, if available.

Section 3.3 *CoPaO An implementation of the InParanoid algorithm*

The new implementation of the InParanoid algorithm was designed to take advantage of new multi-core systems and the HPC that was available. It is important to take advantage of such systems as even for a small number of species the number of comparisons that need to be performed can become quite large – consider the following. If n is the number of species to compare (all-against-all) then there are $(n^2-n)/2$ comparisons that must be performed and $(n^2-n)+n$ BLAST runs. Note that to-date the InParanoid algorithm has only been implemented and extended by one other group. Kim and colleagues created a new InParanoid based algorithm to create “COG” like clusters (Kim, Jung, and Ryu 2006).

This new implementation of the InParanoid algorithm has been called CoPaO (Clusters of Paralogues and Orthologues). Note that all programs are listed relative to the package <org.drd20.bioinformatics>. The CoPaO software suite is composed of the following three core programs: i) a program to distribute CoPaO jobs across an HPC running the Sun Grid engine software <database.copao.CoPaORun>, ii) a program to detect orthologues and paralogues <alignment.copao.CoPaO>, and iii) software to upload the results into an RDBMS <database.copao.CoPaODatabase>. The second program that actually detects orthologues and paralogues can be run standalone – where it is responsible for carrying out the all-by-all BLAST comparisons required by the InParanoid algorithm. This program can only be run on pairs of proteomes. Alternatively the CoPaO HPC software can be used, which accepts a list of species pairs whose proteomes are to be compared. This software distributes the all-against-all BLAST jobs across the HPC before farming out the actual cluster detection to the HPC. The script <database.copao.runOnSelectSpecies.pl> can be used to automate orthologue detection and upload to an RDBMS for a list of species.

Given the large number of comparisons being performed it was decided that the cluster detection program (run after the four BLAST runs) would be threaded (using Perl Interpreter threads). The following tasks carried out by the cluster program have been threaded: i) Running all-against-all BLAST comparisons – only activated when running in standalone mode ii) Sorting of results for each protein into descending order, iii) detection of reciprocal BBHs. The discussion at the end of this chapter addresses the merit of threading the cluster detection program. Note that CoPaO uses a modified version of the BLAST parsing script obtained from the InParanoid authors – which creates a simple table structure of the all-against-all BLAST results.

The InParanoid algorithm has been designed with the assumption that in-paralogues will not diverge from each other at a rate faster than the main orthologues (Remm, Storm, and Sonnhammer 2001). When this assumption is violated in-paralogues will be detected as out-paralogues and thus will not be clustered. This assumption is obviously used to prevent real out-paralogues from being detected as in-paralogues (Remm, Storm, and Sonnhammer 2001).

The theories of neofunctionalization and sub-functionalization have already been discussed. Both methods of in-paralogue divergence may result in two in-paralogues appearing to be more distantly related than a recent gene duplication might suggest. A neofunctionalized in-paralogue is by definition much more likely to have a greater distance to the main orthologue than there is between the two main orthologues.

CoPaO has been designed to attempt to capture such in-paralogues that are further away from the main orthologue than the main orthologues are to each other. They are detected using the following algorithm, which is run after orthologue and in-paralogue detection. For each protein that hasn't been clustered, its intra-genomic BBHs are placed into descending bit score order. If the protein it is most similar to has been clustered it is placed into the same cluster. This process is repeated until no more proteins are added to a cluster. Note that the same BLAST score thresholds are applied to these BBHs as are applied to those used in the previous orthologue/in-paralogue detection routine.

Finally the third program is used to import the CoPaO results into a RDMS.

Section 3.3.1 Confidence value complications

During the first round of in-paralogue detection proteins are only placed into a cluster if they are as close as or closer to the orthologue from the same species than the two orthologues are (one from each species) to each other. Therefore if two orthologues share 100% SI during the first round of paralogue clustering only identical proteins can be clustered with such proteins. The confidence value equation shown in Figure 23 can't be used when two orthologues are identical – because the denominator becomes zero. Note that this occurs because the self-BLAST score of either orthologue will equal the BLAST score between them. Under this situation the confidence value is simply set to 1.

The additional paralogue clustering step incorporates proteins that are, by definition, further away from the main orthologue of the same species than the two orthologues are to each other. Confidence values for these paralogues are always negative. In the rare circumstance that paralogues clustered at this step are placed into a cluster with orthologues that are identical to each other, it is unclear how a confidence value can be assigned. The current version of CoPaO uses the following equation under such circumstances.

$$\frac{Bit(Species[A|B]_{InParalogue}, Species[A|B]_{Orthologue}) - Bit(SpeciesA_{Orthologue}, SpeciesB_{Orthologue})}{Bit(SpeciesA_{Orthologue}, SpeciesB_{Orthologue})}$$

Figure 24: Confidence value equation used for paralogues which are further away from the main orthologue than the orthologues are from each other – when the orthologues are identical.

Section 3.4 Orthologue detection in the PTMDB

In this section a brief description is given of the process used to select species that were to be compared using the CoPaO software. In addition the complication of protein sequence redundancy in the PTMDB is discussed, which complicates the creation of proteome sets for use by the CoPaO software. The method used to remove redundancy from proteome sets is then described along with a brief overview of the estimated completeness of the generated proteome sets.

Section 3.4.1 Species selection

CoPaO uses the graph-based InParanoid algorithm for orthologue/in-paralogue detection. In general graph-based methods are considered faster and easier to automate than tree-based ones (Remm, Storm, and Sonnhammer 2001). That being said the all-against-all BLAST step of the InParanoid algorithm is still quite computationally intensive. The CoPaO software was to be run on a small shared HPC composed of 8 nodes with 16 cores. It was therefore decided that it would only be possible to run the CoPaO software on a limited number of species. As an example of how long it takes to run the CoPaO software on a typical higher Eukaryote, a test was performed running CoPaO in standalone mode on an Amazon EC2 (Elastic Compute Cloud) instance (<http://aws.amazon.com/ec2>). The instance had 7GB of RAM and 8 virtual cores – it took just under 5 hrs to compare the *H. sapiens* and *M. musculus* proteome sets.

Species from each of the three superkingdoms were ordered separately according to the percentage of their proteome that had modification annotations. The top 10 ten species from each superkingdom were then selected from the ordered list. This produced a list of 30 species and 420 comparisons – in comparison version 6.1 of the InParanoid database contained 35 species and 595 comparisons. In addition it was decided that the following recently sequenced species, which were not selected with the above routine, would be compared with *H. sapiens*: *Tetraodon nigroviridis*; *Nematostella vectensis*, and *A. thaliana*. Table 20 shows the full list of species that were selected.

Eukaryotes	Bacteria	Archaea
<i>Mus musculus</i>	<i>Pseudomonas aeruginosa</i>	<i>Sulfolobus acidocaldarius</i>
<i>Tetraodon nigroviridis</i>	<i>Escherichia coli</i> K12	<i>Sulfolobus solfataricus</i>
<i>Bos Taurus</i>	<i>Salmonella typhimurium</i>	<i>Halobacterium salinarum</i>
<i>Rattus norvegicus</i>	<i>Escherichia coli</i> O157:H7	<i>Pyrococcus furiosus</i>
<i>Nematostella vectensis</i>	<i>Escherichia coli</i> O6	<i>Pyrococcus horikoshii</i>
<i>Drosophila melanogaster</i>	<i>Shigella flexneri</i>	<i>Methanothermobacter</i> <i>thermautotrophicus</i> str. Delta H
<i>Gallus gallus</i>	<i>Mycobacterium tuberculosis</i>	<i>Methanocaldococcus jannaschii</i>
<i>Caenorhabditis elegans</i>	<i>Mycobacterium bovis</i>	<i>Mycoplasma pneumonia</i>
<i>Pongo pygmaeus</i>	<i>Bacillus subtilis</i>	<i>Haloferax volcanii</i>
<i>Arabidopsis thaliana</i>		
<i>Schizosaccharomyces pombe</i>		
<i>Saccharomyces cerevisiae</i>		
<i>Homo sapiens</i>		

Table 20: Species selected for orthologue detection using the CoPaO program.

Section 3.4.2 Redundancy in the UniProtKB

The PTMDB contains 72,420 UniProtKB entries for *H. sapiens* which far exceeds the latest estimates regarding the number of protein coding genes of 20,500 (Clamp *et al.* 2007). 53,127 of the *H. sapiens* entries are present in the automatically annotated part of the UniProtKB – TrEMBL. Clamp *et al.* discovered that a great deal of the predicted *H. sapiens* protein coding genes are not evolutionary conserved and suggest that many of these annotations are therefore incorrect (Clamp *et al.* 2007). Note that the TrEMBL database contains all coding sequences from the nucleotide sequence databases EMBL/GenBank/DDBJ. It is therefore likely that some of the annotations that Clamp *et al.* identified as being incorrect have been incorporated into the TrEMBL database and hence the PTMDB. Obviously this problem will extend to other species in TrEMBL.

In addition some entries may contain duplicate sequences, which have accidentally been incorporated into either the Swiss-Prot or TrEMBL databases. Although it is stated that Swiss-Prot entries with the same protein sequence are meant to be merged into a single Swiss-Prot entry (Boeckmann *et al.* 2003). Note that identical sequences from different genes from the same organism are also merged into a single Swiss-Prot entry (UniProtKB 2010). In addition a redundancy removal pipeline is used during the preparation of the TrEMBL database to keep redundancy down to a minimum (O'Donovan *et al.* 1999).

Therefore, before orthologues can be detected between the previously selected species, redundant and erroneous sequences must first be removed from their respective sequence sets in the PTMDB.

Note that the PTMDB contains 6,074,524 UniProtKB entries - 366,226 Swiss-Prot and 5,708,298 from TrEMBL.

Section 3.4.3 Removing redundancy from the PTMDB

UniProtKB is annotated with what are called 'complete proteome sets', currently for 1,380 species. In addition the Integr8 (Pruess, Kersey, and Apweiler 2005) project provides similar proteome sets, currently for 1,515 species. At the time these sets were being constructed there were discrepancies between the sizes

of some of the proteomes and their respective protein encoding-gene counts. For example the *H. sapiens* Integr8 set contained 40,639 UniProtKB accessions – far more than the 20,500 protein encoding genes predicted by Clamp *et al.* (Clamp *et al.* 2007). Additionally note that some proteins that are present in the PTMDB may be missing from Integr8 dataset (and visa-versa); note that this point does not apply to the ‘complete proteome sets’ as these annotations are taken from the same version of UniProtKB that the PTMDB was populated with. It was therefore decided that, where available, these proteome sets would provide a starting point for further redundancy removal.

The protocol that is described next, to create proteome sets for the PTMDB, is in-part based on that used by the Integr8 project to create their *H. sapiens* proteome set (see Pruess *et al.*, 2005) (see Figure 25 for an overview of this protocol).

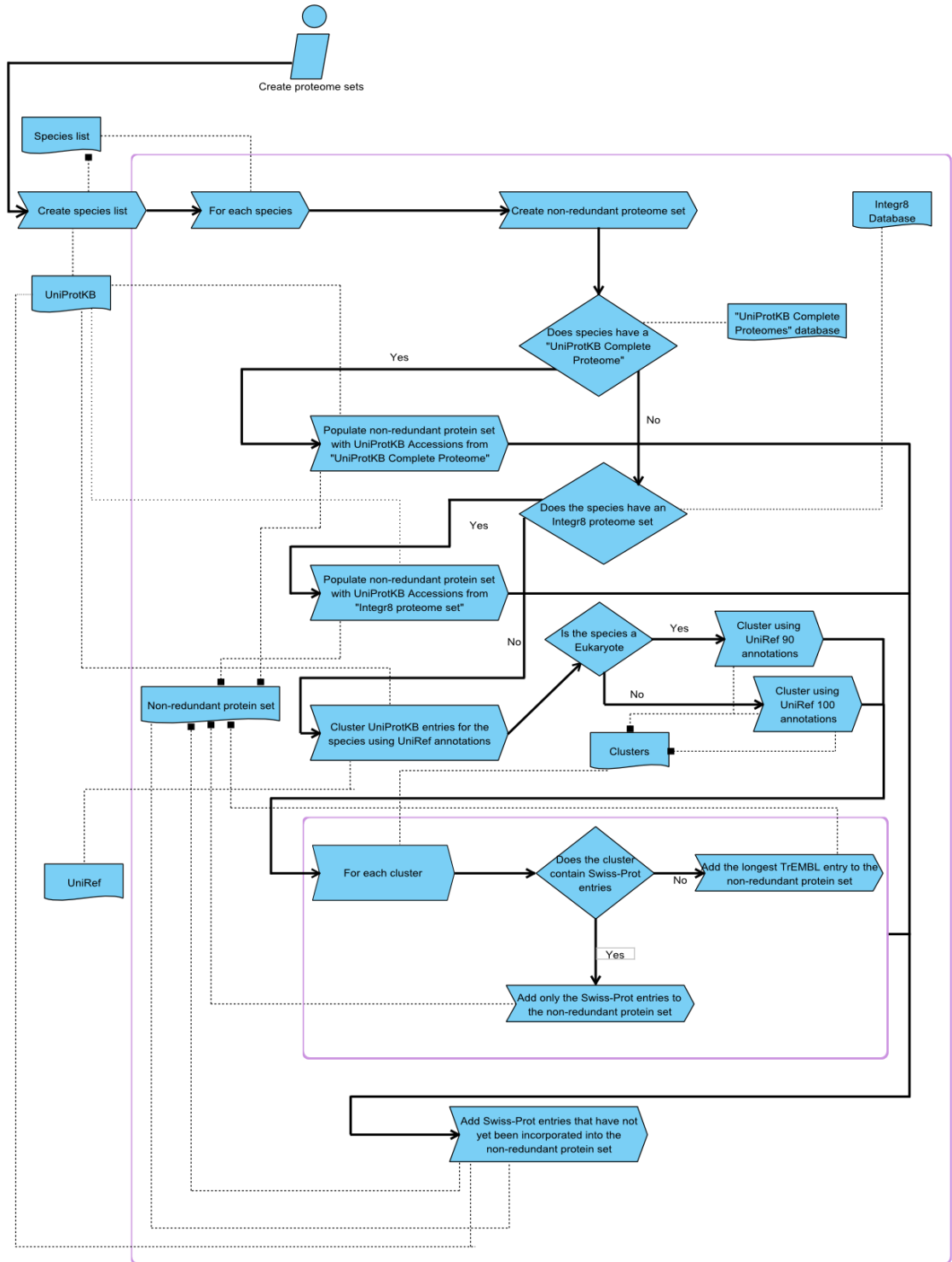
Section 3.4.3(a) Step one – Initial selection of proteins

Step one of this protocol creates an initial set of proteins for each species – according to the following two rules:

- I. All Swiss-Prot and TrEMBL entries are extracted for a species that isn't in either the UniProtKB Complete Proteomes or Integr8 project
- II. Alternatively only those entries that are included in the proteome set are extracted. Note that when a species is present in both projects the UniProtKB Complete Proteomes set is taken, as this has been expertly curated.

Section 3.4.3(b) Step two – UniRef clusters are overlaid

Step two involves the use of UniRef (Suzek *et al.* 2007) clusters to identify redundant sequences. The UniRef project provides non-redundant UniProt protein reference clusters at the predefined sequence identity thresholds of 100, 90 and 50 PSI (Percentage Sequences Intity) (Suzek *et al.* 2007). The UniRef database is seeded with the primary protein sequences listed for each UniProtKB entry as well as those splice variant sequences represented in the feature table of each entry (Suzek *et al.* 2007).



Key



Data/output



Process



Data moving into process



Data moving out of process



For each loop

Figure 25: Redundancy removal protocol.

Proteins present in each proteome were clustered by their corresponding UniRef cluster accessions (thereby grouping similar proteins together). It is possible for proteins to be present in multiple clusters by virtue of the differential clustering of their corresponding splice-variants; such overlapping clusters are automatically merged.

It was assumed that there would be more duplicate and/or sub-fragment sequences present in TrEMBL for Eukaryote species compared with Prokaryote ones. For this reason, protein sets for Eukaryote species were clustered at the more restrictive 90 PSI compared with 100 PSI used for Prokaryotes. Obviously the use of the 90 PSI threshold for Eukaryote species increases the likelihood of excluding sequences that arise from unique genes.

Section 3.4.3(c) *Step three – redundancy removal*

Now that the proteins in each proteome have been clustered, redundant sequences are identified as follows, and subsequently removed from the corresponding proteome. All proteins present in the Swiss-Prot section of UniProtKB are protected and never deleted. All TrEMBL entries that have been clustered with at least one Swiss-Prot entry are first removed from the proteome (because the gene they represent has presumably already been accounted for by the Swiss-Prot entry). Finally the longest TrEMBL entry is kept (and all others removed) from clusters which do not contain any entries from the Swiss-Prot database.

Section 3.4.3(d) *Species selection*

As will become apparent in the next chapter, proteome sets were required for more than just orthologue detection. Therefore proteome sets have been created for all species in the PTMDB that have an Integr8 or UniProtKB Complete Proteomes set. In addition, species present in UniProtKB were grouped by super-kingdom and ranked in descending order according to how many entries they had – those in the top 10 of each list were also added to the list of species for which proteomes should be generated. This same process was carried out for the PTMDB – ranking species by the number of PTM annotations they had. Table 21 provides a breakdown of the number and source of species that had been selected for proteome generation. In total

1472 species were selected for proteome generation: 72 Eukaryotes, 54 Archaea, 582 Bacteria and 763 taxa that do not belong to any of the super-kingdoms.

Database	Total	Eukaryotes	Archaea	Bacteria	Other
UniProtKB CP	947	42	48	543	314
Integr8	551	2	5	60	461
PTMDB	4	1	3	-	-
Swiss-Prot	1	1	-	-	-
UniProtKB	9	3	1	5	-
Sub-Total	1512	72	57	608	775
Not in UniRef	40	-	2	26	12
Final-Total	1472	72	55	582	763

Table 21: Number of species for which proteome sets have been generated subdivided by super-kingdom (note that other contains taxa that don't belong to a super-kingdom – e.g. Viruses). The number of proteomes imported from the UniProtKB Complete Proteomes and Integr8 projects are shown. In addition the number of species selected for proteome generation, by being in the top 10 most abundant species (in their corresponding super-kingdom) in the PTMDB, Swiss-Prot, and UniProtKB, are shown (note this excludes those already selected by being present in the UniProtKB Complete Proteomes or Integr8 project). The number of species that had to be removed because they didn't have any entries in the UniRef database is also shown.

Section 3.4.3(e) Redundancy removal summary

Table 22 displays the number of proteins present after each stage of the proteome construction process. At the start of the process, for the 1,472 species - there were 3,236,891 UniProtKB entries. All UniProtKB entries belonging to species which were not present in their corresponding UniProtKB Complete Proteome Integr8 proteome set were removed (obviously this doesn't apply to species that don't have an entry in either). This step removed 3,636 Swiss-Prot and 340,923 TrEMBL entries – 80% of these were assigned to Eukaryote species. Proteins were next removed that did not have UniRef annotations – a requirement for the final step. This step removed 344,559 proteins in total – 86% of these were removed from bacterial species. Only 72 of these entries were contributed from the Swiss-Prot database. The final step was designed to remove any remaining redundancy that existed in the proteomes constructed so far. This procedure involved the removal of proteins contributed from TrEMBL, which clustered with those from Swiss-Prot. In addition, only the longest TrEMBL entry was retained from a cluster of TrEMBL-only proteins. 90,605 TrEMBL entries were removed. 84,511 of these were from Eukaryote species – proteins of which were clustered at 90% SI. Only 5,655 proteins were removed at this step for bacterial species.

The final set of proteomes contained 2,768,148 proteins; 68% from Bacteria, 27% from Eukaryotes and the remaining 5% is distributed between Archaea and those species that do not belong to any of the superkingdoms.

UniProtKB (A)			
	Swiss-Prot	TrEMBL	Total
Eukaryotes	92695	1010650	1103345
Bacteria	191438	1794545	1985983
Archaea	13601	97960	111561
Other	3314	32688	36002
Total	301048	2935843	3236891

Overlay of Integr8 and UniProtKB CP (B)			
	Swiss-Prot	TrEMBL	Total
Eukaryotes	90065	736585	826650
Bacteria	190583	1732452	1923035
Archaea	13589	97390	110979
Other	3175	28493	31668
Total	297412	2594920	2892332

UniRef constraint overlay(C)			
	Swiss-Prot	TrEMBL	Total
Eukaryotes	92695	1010650	1103345
Bacteria	191438	1794545	1985983
Archaea	13601	97960	111561
Other	3314	32688	36002
Total	301048	2935843	3236891

Final Proteomes (D)			
	Swiss-Prot	TrEMBL	Total
Eukaryotes	90028	647667	737695
Bacteria	190557	1697746	1888303
Archaea	13589	97265	110854
None	3166	28130	31296
Total	297340	2470808	2768148

Reduction			
	Swiss-Prot	TrEMBL	Total
	2630	274065	276695
	855	62093	62948
	12	570	582
	139	4195	4334
	3636	340923	344559

Reduction			
	Swiss-Prot	TrEMBL	Total
	37	4407	4444
	26	29051	29077
	0	0	0
	9	4	13
	72	33462	33534

Reduction			
	Swiss-Prot	TrEMBL	Total
	0	84511	84511
	0	5655	5655
	0	125	125
	0	359	359
	0	90650	90650

Table 22: Number of proteins removed at each step of the redundancy removal protocol.

Section 3.4.3(f) *Estimated proteome completeness*

In order to estimate the completeness of the proteome sets just created - proteome sizes for selected model organisms have been compared with the number of protein-coding genes listed in the Ensembl (Flicek *et al.* 2010) and CMR (Comprehensive Microbial Resource) (Davidsen *et al.* 2010) databases for Eukaryotes and Prokaryotes, respectively.

Eukaryotes

Table 23 shows the estimated proteome coverage for selected model eukaryotes. The proteome for *T. nigroviridis* appears at the top of Table 23 as its percentage coverage value is the highest. This proteome contains 75% more proteins than there are protein coding genes in Ensembl for this species.

For simpler eukaryotes, such as *Caenorhabditis elegans*, it is therefore almost certain that the Ensembl dataset is far from complete. The *H. sapiens* proteome contains 28,457 proteins – higher than the number of Ensembl protein coding genes (23,438).

Table 23 shows that there is high proteome coverage for the following other eukaryotes: *Drosophila melanogaster*, *M. musculus*, *N. vectensis*, *Saccharomyces cerevisiae*, *C. elegans*, *Schizosaccharomyces pombe*. There is also greater than 50% coverage for: *Danio rerio*; *Bos taurus*, *Rattus norvegicus*. The two plants in this table: *A. thaliana*; *Oryza sativa Indica* aGroup, also have high proteome coverage.

PTMDB					
Species	Swiss-Prot UniProtKB	Proteome size rel. 13.3	Genome sequence statistics Number of protein coding genes	Source	PTMDB coverage
Tetraodon nigroviridis	46	26,234	15,204	Ensembl	172.55%
Homo sapiens	19,293	28,457	23,438	Ensembl	121.41%
Drosophila melanogaster	2,784	15,303	14,076	Ensembl	108.72%
Arabidopsis thaliana	6,582	26,723	25,498	1	104.80%
Mus musculus	15,447	24,038	22,974	Ensembl	104.63%
Nematostella vectensis	28	22,007	22,000	2	100.03%
Saccharomyces cerevisiae	6,555	6,526	6,532	Ensembl	99.91%
Caenorhabditis elegans	3,178	20,002	20,158	Ensembl	99.23%
Schizosaccharomyces pombe	4,217	4,942	4,985	Ensembl	99.14%
Oryza sativa Indica Group	423	36,988	46,022-55,615	3	80.37%
Danio rerio	2,078	16,102	24,147	Ensembl	66.68%
Bos Taurus	5,180	11,483	20,118	Ensembl	57.08%
Rattus norvegicus	6,979	11,167	21,107	Ensembl	52.91%
			26,835-22,000	4	50.76%

Table 23: Example estimated proteome coverage in the PTMDB proteome sets. The resources that the number of protein coding genes was obtained from 1[^](Initiative 2000), 2[^](Putnam et al. 2007), 3[^](Yu et al. 2002), 4[^](Elsik, Tellam, Worley, et al., with Sequencing et al., Analysis Consortium 2009).

Prokaryotes

Estimated proteome coverage in the PTMDB for selected model Eubacteria and Archaea is shown in Table 24. For most of the selected species, the majority of their proteomes appear to be in the TrEMBL database. The size of almost all of

the shown proteomes is very close to the number of protein coding genes listed in the CMR database.

Species	PTMDB		Estimators		
	Swiss -Prot	Proteome size	Genome sequence statistics		
	UniProtKB rel. 13.3		Number of protein coding genes	Source	PTMDB coverage
Eubacteria					
Streptococcus pneumonia	504	2106	2,047	CMR	102.88%
Streptomyces coelicolor	698	8029	7,897	CMR	101.67%
Escherichia coli K12	4343	4339	4,289	CMR	101.17%
Helicobacter pylori	578	1553	1,544	CMR	100.58%
Bacillus subtilis	2871	4105	4100	CMR	100.12%
Mycoplasma pneumoniae	687	687	688	CMR	99.85%
Corynebacterium diphtheria	311	2265	2,320	CMR	97.63%
Clostridium difficile 630	223	3704	3,804	CMR	97.37%
Enterococcus faecalis	442	3235	3,337	CMR	96.94%
Mycobacterium tuberculosis	1439	3943	4,246	CMR	92.86%
Escherichia coli	616	7943			
Archaea					
Pyrococcus horikoshii	452	2076	2,064	CMR	100.58%
Sulfolobus tokodaii	343	2816	2,826	CMR	99.65%
Sulfolobus acidocaldarius	299	2206	2223	CMR	99.24%
Pyrococcus furiosus	429	2043	2,065	CMR	98.93%
Methanosarcina mazei	443	3298	3,381	CMR	97.55%
Haloquadratum walsbyi DSM 16790	153	2637	2,862	CMR	92.14%

Table 24: Example proteome coverage in the PTMDB for selected model species and species of special interest.

Section 3.5 Results

Section 3.5.1 Cluster Validation

Section 3.5.1(a) Indirect cluster comparison

The orthologue assignments created by the InParanoid 6.1 software between *H. sapiens* and *M. musculus* were downloaded from the InParanoid web site. In addition a FASTA file for each species was also downloaded from the same website. These files identified protein sequences with identifiers from Ensembl for *H. sapiens* and MGI (Mouse Genomics Initiative) for *M. musculus*. To compare the two assignments, the Ensembl and MGI identifiers needed to be

mapped to their corresponding UniProtKB identifiers in the PTMDB. This was accomplished by simply creating one table that contained the checksum of all UniProtKB sequences in the PTMDB and another, which contained the checksum of those in the two FASTA files obtained from the InParanoid web site. An intersection of these two tables was then obtained – using the checksum field. This method is of course susceptible to minor differences in the protein sequences stored in these two different datasets. Table 25 shows that only 65% of *H. sapiens* and 75% of *M. musculus* proteins from the InParanoid dataset could be mapped into the PTMDB.

Species	Number of proteins	
	InParanoid	PTMDB
<i>Homo sapiens</i>	22983	15055
<i>Mus musculus</i>	23132	17498
Total	46115	32553

Table 25: The number of *Homo sapiens* and *Mus musculus* proteins in the InParanoid 6.1 dataset and the corresponding number that could be mapped to UniProtKB entries in the PTMDB.

The orthologue assignments were then compared between the CoPaO and InParanoid datasets. A new dataset was first created that contained all the clusters from the InParanoid dataset, but excluded any proteins that had not been mapped to the PTMDB. Next any cluster that did not contain at least one protein from *H. sapiens* and *M. musculus* was also thrown away. Finally the percentage of proteins that had been placed into the same cluster of both datasets was deduced – according to the following process. A protein was said to be in the same cluster in the CoPaO dataset as in the InParanoid dataset – if its CoPaO cluster ID had been assigned to the majority of the other members of the same InParanoid cluster (note that the cluster ID was set to -1 for proteins not in a CoPaO cluster). Of the 32,523 proteins from the InParanoid dataset that were mapped into the PTMDB – 18,146 were present in the constrained InParanoid cluster set. 16,745 of these were determined to be in the same cluster in both the InParanoid and CoPaO datasets.

Section 3.5.1(b) Gene Ontology Analysis

Homology refers only to common ancestry and not to the degree of conservation between two sequences (Koonin 2005). Also note that orthology

detected between proteins does not imply that they share the same function; although they commonly do (Koonin 2005). Therefore one method of validating the correct detection of orthologous proteins is to determine whether the orthologues share the same function, noting the caveat above. Such a technique will inevitably be dependent on the quality and completeness of the underlying functional annotations.

The Gene Ontology (GO) is the gold standard of protein function annotation (Ashburner *et al.* 2000). The GO was created to standardise the vocabulary that is used to describe the product(s) of a gene. The Gene Ontology is composed of three namespaces: Molecular Function; Cellular Component and Biological Process. The process to which a gene product contributes to is described by using the Biological Process ontology, e.g Cell Growth, Translation, DNA Replication etc. The role that a gene product plays in a biological process is more precisely described by the Molecular Function ontology, e.g catalyses binding, protein binding, GTPase activating protein binding, transporter etc. The location of the active gene product within the cell is described using the Cellular Component ontology, e.g Nucleus, Mitochondria etc.

The Gene Ontology allows a single term to be connected to many parents. This structure is referred to as a DAG (Directed Acyclic Graph). In the Gene Ontology there are two different types of edges that can be used to connect terms together, these are: *is_a* and *part_of*. An *is_a* connection defines an abstract relationship between the parent and child term. This relationship resembles that from Object Orientated programming where a subclass inherits all attributes and methods from its parent class. A *part_of* relationship defines that when the child term is present it is always part of the parent term. Both types of relationship follow the 'true path' rule. This rule states that not only should a specific annotation apply to a gene, but also that all parents of this annotation should apply to the gene as well.

The simplest validation procedure would only compare whether proteins in the same cluster shared identical functional annotations. However the possibility of subfunctionalization and neofunctionalization having occurred to proteins in a

cluster makes it less than desirable to assume that functional annotations should be identical.

A validation procedure is now presented which scores the functional distance observed between orthologues and in-paralogues in the same cluster. This procedure uses a metric, which provides a score indicating the semantic similarity of two GO terms. Note that GO semantic similarity measures have previously been shown to have good correlation with sequence similarity (Lord *et al.* 2003). The RSS (Relative Specificity Similarity) semantic metric published by Wu *et al.* 2006 was chosen as it scores GO term pairs according to both the distance between the terms and the specificity of each term (Wu *et al.* 2006). GO terms that represent general abstractions of a gene function are found higher up in the DAG (Ashburner *et al.* 2000).

The RSS metric assigns a value of one to GO terms that are identical and are the most specific in the whole of the DAG. The most specific terms in the DAG are taken as those that are the maximum distance from the root of node of the ontology (i.e. the deepest). The remaining possible GO term pairs are assigned a value less than one and greater than or equal to zero. A pair of GO terms is assigned a score of zero if their most recent common ancestor is the root node of the DAG.

Lord *et al.*, 2003 previously demonstrated that the correlation between GO semantic similarity measures and sequence similarity was greatest when using the molecular function namespace. The validation routine was therefore setup to use this namespace – although initial results produced by this routine were similar regardless of which namespace was used (data not shown).

To utilise the RSS metric in the validation routine, GO annotations were first obtained from the GOA (Gene Ontology Annotation) resource (Barrell *et al.* 2009). The gene ontology is available for download in OBO (Open Biomedical Ontology) format – a Java parser and object representation was created for version 1.2 of the OBO specification (See <<http://www.geneontology.org/GO.format.obo-1.2.shtml>> for the specification). Next the RSS metric was implemented in the same software package as the new parser.

It was decided that only species comparisons that involved *H. sapiens* as one of the two species would be validated with the RSS metric. For each cluster an average RSS value was first calculated. An all-by-all comparison was performed between proteins in the same cluster which have at least one GO term. As originally published by Wu *et al.*, (2006), when multiple GO term annotations are present for either, or both, proteins being compared an all-by-all comparison of the GO terms is performed. The maximum RSS value is taken to represent the semantic similarity of the two proteins. An average is then calculated for all RSS values that have been calculated. Note that this process does not take into account the percentage of the Proteins in the clusters that actually had GO annotations.

To provide a context for the intra-cluster average RSS values, inter-cluster average RSS values have been calculated. These were produced by randomly selecting 1000 clusters, or the maximum number of clusters present, whichever was greater. Each selected cluster was then compared to all others. Not all clusters have one protein from each species being compared annotated with a GO term. Since such clusters have not been compared; the percentage of clusters that could be compared has been recorded and is shown in Table 26.

The validation procedure shows that proteins in the same CoPaO cluster have on average a smaller functional distance than proteins in different clusters. The intra-cluster average value varies from 1 – 0.75 between the *H. sapiens* proteome and those of the listed species. Whereas the inter-cluster value ranges from 0.73-0.45, with a value of 1 associated with two annotations at the root of the gene ontology graph.

The average intra-cluster distance between *H. sapiens* and the prokaryotes listed in Table 26 is slightly smaller than that observed for the eukaryote comparisons. Presumably this has occurred because the only proteins retained in both eukaryotes and prokaryotes have essential functions, which are unlikely to change over the course of evolution.

Average RSS			
Species	Intra-cluster	Inter-cluster	Percentage of clusters compared
Tetraodon nigroviridis	1.00	0.44	33.00%
Pyrococcus horikoshii	0.99	0.65	100.00%
Methanocaldococcus jannaschii	0.98	0.63	100.00%
Salmonella typhimurium	0.97	0.62	100.00%
Mycobacterium tuberculosis	0.96	0.67	100.00%
Mycobacterium bovis	0.96	0.65	100.00%
Shigella flexneri	0.96	0.61	89.00%
Arabidopsis thaliana	0.96	0.55	89.00%
Methanothermobacter thermautotrophicus str. Delta H	0.95	0.71	89.00%
Pyrococcus furiosus	0.95	0.52	100.00%
Saccharomyces cerevisiae	0.95	0.55	89.00%
Mycoplasma pneumoniae	0.94	0.65	89.00%
Pseudomonas aeruginosa	0.94	0.70	67.00%
Escherichia coli K12	0.94	0.69	100.00%
Sulfolobus solfataricus	0.94	0.69	100.00%
Escherichia coli O157:H7	0.93	0.61	100.00%
Schizosaccharomyces pombe	0.93	0.52	78.00%
Pongo pygmaeus	0.92	0.51	78.00%
Drosophila melanogaster	0.92	0.46	44.00%
Bacillus subtilis	0.91	0.67	100.00%
Halobacterium salinarum	0.91	0.73	100.00%
Escherichia coli O6	0.90	0.65	100.00%
Rattus norvegicus	0.89	0.56	89.00%
Mus musculus	0.87	0.54	67.00%
Bos Taurus	0.85	0.51	33.00%
Gallus gallus	0.84	0.50	78.00%
Sulfolobus acidocaldarius	0.83	0.68	100.00%
Caenorhabditis elegans	0.82	0.49	89.00%
Nematostella vectensis	0.75	0.45	78.00%

Table 26: RSS validation of orthologue clusters detected between Homo sapiens and a select list of other species.

The percentage of CoPaO clusters for which a functional distance could be calculated was on average higher for *H. sapiens*/prokaryote comparisons than *H. sapiens*/eukaryote ones. It is likely that this results from a combination of: a) better annotation of prokaryote proteomes, which tend to be smaller, and b) that those orthologues detected carry out essential functions, which may be more likely to be annotated with a GO term.

Section 3.5.2 Proteome conservation

One implicit assumption, which is formed from the definition of homology, is that species that diverged from each other more recently should be more homologous to each other than those that diverged more distantly. Table 27 shows the degree of homology detected between *H. sapiens* and a collection of model organisms – using the CoPaO software. In this table species have been ordered according to the percentage of the *H. sapiens* proteome which has been annotated as an orthologue in each comparison.

Note that “percentage orthology” can be expressed from the point of view of either species being compared. Given a comparison between two species the number of orthology assignments must be equal for both species by definition; the same is not true for the proteome size of the two species being compared. This statement is only violated when multiple in-paralogues are 100% identical and thus are all assigned as iso-orthologues. Table 27 shows “percentage orthology” from the point of view of *H. sapiens* and the species with which its proteome is being compared to.

This table also shows the percentage of each proteome that has been assigned as either an orthologue or in-paralogue. Species that have an incomplete proteome are highlighted.

This table shows that the most orthologous species to *H. sapiens* is *M. musculus*. 57% of the *H. sapiens* proteome is orthologous to the *M. musculus* proteome (which is 68% orthologous to the *H. sapiens* proteome). ~5% of the *H. sapiens* proteome has been assigned as an in-paralogue. *T. nigroviridis* follows *M. musculus*. 34% of the *H. sapiens* proteome is orthologous to the *T. nigroviridis* proteome (which is 36% orthologous). These values suggest that

H. sapiens and *T. nigroviridis* have undergone significant proteome evolution since they diverged.

B. taurus, *R. norvegicus*, *Gallus gallus*, and *Pong pygmaeus* all diverged from *H. sapiens* later than *T. nigroviridis* but appear below it in the table. These have all been highlighted in this table – because their proteomes are currently incomplete in the PTMDB. A higher percentage than for *M. musculus* of each of these species' proteomes has been assigned an orthologous relationship. These results can probably be attributed to a combination of the following: a) essential or genes with disease links are given sequencing/annotation priority, and b) genes that share a high percentage identity are annotated first as their assignments are relatively unambiguous.

The next species with complete proteomes in this table are *N. vectensis* and *D. melanogaster*. Note that although the *H. sapiens* proteome appears closer to *N. vectensis*, the percentage of the *D. melanogaster* proteome that is assigned an orthologous relationship is higher. The *C. elegans* proteome is 20% orthologous to that of *H. sapiens*. Of all the Eukaryote species shown in this table, the *H. sapiens* proteome is least orthologous to that of the two yeast species shown. However, the percentage of the yeast species proteomes that are orthologous to that of *H. sapiens* is relatively high. For example 41% of the *S. pombe* proteome has been assigned an orthologous relationship. This may suggest that approximately half of the yeast proteome is essential in higher Eukaryotes such as *H. sapiens*.

This table shows that only ~1% of the *H. sapiens* proteome is assigned as being orthologous to proteomes from bacteria. However ~10% of each bacterial proteome has been found to be orthologous to the *H. sapiens* proteome. Again the most plausible explanation is that the proteins represented in the 10% are essential across both super-kingdoms.

	Percentage of proteome with an orthologous partner		Percentage of proteome clustered	
	Homo sapiens	Species X	Homo sapiens	Species X
Mus musculus	57.47	68.05	62.48	72.57
Tetraodon nigroviridis	34.01	36.82	45.97	39.47
Bos Taurus	33.29	82.73	36.96	84.87
Rattus norvegicus	31.36	79.79	36.08	83.14
Nematostella vectensis	19.8	25.38	33.74	31.5
Drosophila melanogaster	17.16	31.8	31.2	39.85
Gallus gallus	14.91	76.92	19.25	80.67
Caenorhabditis elegans	14.18	20.04	28.82	25.33
Pongo Pygmaeus	12.74	83.46	14.69	84.03
Arabidopsis thaliana	9.78	10.58	19.33	27.02
Schizosaccharomyces pombe	7.27	41.3	12.71	45.29
Saccharomyces cerevisiae	6.44	27.9	13.57	32.72
Pseudomonas aeruginosa	2	10.16	4.47	13.75
Escherichia coli K12	1.76	11.27	3.65	14.31
Salmonella typhimurium	1.73	10.74	3.94	13.46
Escherichia coli O157:H7	1.71	9.12	3.95	11.61
Escherichia coli O6	1.71	9.03	3.75	11.58
Shigella flexneri	1.61	11.09	3.43	13.51
Mycobacterium tuberculosis	1.55	11.06	3.72	14.58
Mycobacterium bovis	1.55	11.24	3.66	14.98
Bacillus subtilis	1.54	10.6	3.94	14.64
Sulfolobus acidocaldarius	1.34	16.55	3.2	20.44
Sulfolobus solfataricus	1.31	12.58	2.69	16.44
Halobacterium salinarum	1.24	14.34	2.82	17.53
Pyrococcus furiosus	1.2	16.4	2.39	19.97
Pyrococcus horikoshii	1.14	15.17	2.34	18.11
Methanothermobacter thermautotrophicus str. Delta H	1.08	16.21	2.12	18.78
Methanocaldococcus jannaschii	0.94	14.65	1.99	16.61
Mycoplasma pneumonia	0.46	19.07	1.27	21.25
Haloferax volcanii	0.11	28.57	0.83	32.14

Table 27: Percentage of the Homo sapiens proteome which has been clustered with a selected group of species.

Section 3.6 Discussion

Orthologue annotations have successfully been added to the PTMDB with a new implementation of the InParanoid algorithm. 420 pairs of species were analysed – composed of 30 individual species. Orthologue assignments between *H. sapiens* and *M. musculus* stored in the InParanoid database (version 6.1) have been compared with those produced by the CoPaO software. 65% of the *H. sapiens* and 75% *M. musculus* proteins in the original InParanoid dataset were mapped into the PTMDB. The same sequence was identified in each dataset by an MD5 checksum. This precludes the detection of sequences which represent the same gene in each database that have any differences at all in their sequence. The alternatives would have been to either: 1) run an exhaustive all-against-all sequence comparison between the two datasets, or 2) attempt to use an ID mapping service to between the MGI/Ensembl and the UniProtKB. Regardless the majority of the InParanoid dataset was mapped into the PTMDB. It was discovered that 92% of the orthologue assignments in the InParanoid dataset (which could be compared) matched those produced by the CoPaO software. Additionally further validation of the orthologous clusters has been provided through the use of the GO semantic similarity measure.

An error was discovered in the CoPaO implementation of the InParanoid algorithm that resulted in a small number of orthologues also being assigned to different clusters as in-paralogues. 79,751 paralogues (including both first and second pass detection) of these 2,617 were found to be incorrectly annotated; they were in fact correctly annotated as orthologues in different clusters. These entries have been removed from the clusters in which they were present.

After the redundancy-removal step, the *H. sapiens* proteome size far exceeded that of the number of protein coding genes predicted by Clamp *et al.*, (2007). Although it was suspected that redundant sequences of some form existed in this set, nothing was done to attempt to reduce this protein set any further. An analysis has now been performed on the *H. sapiens* and *M. musculus* proteome set with an interesting finding. A checksum was created for all sequences present in the PTMDB for both of these species. Identical checksums were then identified. Surprisingly 3,784 Swiss-Prot and 723 TrEMBL entries were

found to have duplicate sequences for *H. sapiens*. A similar number was seen for *M. musculus* with 4,068 Swiss-Prot and 40 TrEMBL entries identified. Unfortunately Swiss-Prot entries were explicitly excluded from being removable during the redundancy removal pipeline. It is interesting to note that two of the duplicate entries A8K5L3 and P05109 have been merged into a single entry in Swiss-Prot version 56.7. In the future when the PTMDB is updated, Swiss-Prot entries identified as being identical will have their PTM annotations merged into a single entry.

The program `<org.drd20.bioinformatics.alignment.CoPaO>` makes use of Perl interpreter threads in an attempt to speed up the detection of orthologues and paralogues from the BLAST all-by-all comparisons. The design of this Perl program made it difficult to test the merits of threading the cluster detection; therefore a Java version of this program has been created `<org.drd20.bioinformatics.alignment.CoPaO.java>`. On an Intel Core II (2.4GHZ) with 4GB RAM the cluster routine completed its detection of orthologues between *H. sapiens* and *M. musculus* 16% faster when running two threads compared to one (a thread count greater than two became progressively slower). Note that the Perl program never actually completed this process on the benchmark machine – due to an unidentified bug (which appeared to be related to Perl itself), not releasing memory after a thread had finished (this behaviour was observed regardless of whether Perl was compiled against the system malloc or its own). The Java program was able to carry out all 420 comparisons (minus the BLAST all-by-all comparisons) in ~30mins; therefore it appears that there is little to gain by distributing the orthologue detection around a compute-cluster (obviously this still depends on the number of species being compared.); however it is still very much advisable to distribute the BLAST all-by-all comparisons on such a cluster. With only 30 species and few with large proteomes it also appears that there is little to gain by threading the orthologue detection; although it should be pointed out that for the Java program this was a trivial process.

Chapter 4

Cross annotation of PTMs

Section 4.1 *Summary*

To obtain accurate conclusions regarding the conservation of post-translational modifications between species, and specifically between orthologues and paralogues, requires that the modification dataset used must be as complete as possible. A preliminary analysis suggested that numerous plausible modification sites had not yet been annotated. In addition, many of the proteome sets that were created in the last chapter contained a significant number of TrEMBL entries for which no annotations existed in the PTMDB. Hence the aim of the work presented in this chapter was to develop a method to transfer annotated acceptor sites between homologous sites. With particular emphasis placed on extending the PTMDB into the sequence space occupied by entries in the TrEMBL database.

This chapter begins with an explanation of why there has been a massive explosion in the number of proteins for which their modification states need to be identified. A brief description then follows on the high and low throughput experimental techniques that can be used to identify PTMs. This is followed by an overview of the computational techniques that are used to support and extend the data that is produced by experimentalists. Particular emphasis is placed on the continued requirement to design tools that can predict PTMs and transfer annotations between homologous sites. A protocol is then described that has been used to cross-annotate PTM annotations between proteins in Swiss-Prot and TrEMBL. Finally a summary of the cross-annotations that have been made is provided.

Section 4.2 *Introduction*

Next generation sequencing technologies (Metzker 2010) and the knowledge and expertise gained from sequencing the Human genome (Lander *et al.* 2001; Venter *et al.* 2001) have resulted in an explosion in the number of genome sequences that have become available. The Genomes OnLine Database (GOLD) (Liolios *et al.* 2010), which documents active and on-going genome sequencing projects, has seen an explosion in the number of projects referenced. The authors of GOLD noted that in September 2005 there were 1,575 projects referenced, which had expanded to 5,843 in September of 2009.

At the same time as the publication of the draft Human genome sequence, the HUPO (Human Proteomics Organisation) was announced, with the goal of supporting researchers in the exploration of the Human proteome (seen as the next target) (Abbott 2001). A major component of the process of defining and analysing a proteome is to discover when (i.e. specific biological context) and where (defined at either the protein or residue level) PTMs occur.

Section 4.2.1 *Experimental determination*

An overview of the methods that can be used to identify PTMs can be seen in Figure 5. With so many genomes now sequenced (and many more to follow) the sheer number of proteins whose modification patterns (i.e. when and where) need to be discovered quickly overwhelms the traditional low-throughput PTM discovery techniques (Gupta *et al.* 2007).

One such low-throughput technique involves identifying a protein of interest that appears to shift position on 1D and 2D SDS PAGE gels – where one gel contains proteins extracted from a sample where an inhibitor for a specific modification has been added indicating a gain/loss, or rearrangement, of the chemical structure of the protein. Modifications may be prevented by specific enzyme inhibitors (Davis *et al.* 2006), or by the removal of a gene which is required for the modification to take place. In addition, modifications can be mapped to specific residues using site directed mutagenesis in combination with an identification technique such as 2D SDS PAGE (Robinson and Michel 1995). This technique is considered low-throughput because of a) the

requirement to be able to identify the protein represented by a gel band (i.e. requires an antibody that recognises a protein of interest or the use of Edman sequencing), and b) if the modified residue needs to be identified – this may require repeated site-directed mutagenesis experiments.

The inability rapidly to identify the proteins that are present in each gel band has recently been overcome by the development of advanced analysis pipelines that use mass spectrometers for protein identification. Not only do they allow for protein identification but they can also be used to elucidate the position and nature of any PTMs that may be present, removing the requirement for site-directed mutagenesis (Claverol *et al.* 2003). Even with the availability of high-throughput techniques, those that are low-throughput are still routinely used to validate results from the aforementioned techniques.

Section 4.2.2 Computational determination

Figure 5 shows that a significant number of databases and tools have been designed to aid in the identification of PTMs. These bioinformatics resources can be broken down into three groups: i) databases, ii) tools used to aid in the interpretation of mass spectrometry results, and iii) those that are used to predict PTM sites. Databases that contain both experimentally determined and predicted PTM annotations have many uses during the process of proteome characterisation. One example is their use in identifying sites that should be mutated for PTM mapping – via annotated homologous sites. Tools that are used to detect PTMs in mass spectrometry results also make use of PTM databases. For example, in the process of annotating mass spectra with glycan structures, the GlycoWorkbench tool allows users to browse known structures in the CarbBank and CFG (Consortium for Functional Glycomics) databases (Ceroni *et al.* 2008).

The design of tools to predict PTM sites predates the use of high-throughput experimental proteomic techniques. However such tools present many useful attributes leading to their continued development. Most PTM prediction tools do not attempt to deduce the biological context that is required for a PTM to occur. This can be seen as both a positive and negative aspect of prediction tools; although most cannot tell experimentalists the conditions they require to

observe a PTM, they can at least tell them that conditions may exist that will trigger the modification of a particular residue. This is in sharp contrast to experimental approaches that can only be used to identify PTMs which are actually present in their system under very precise conditions. Some PTM prediction tools attempt to predict some aspects of the biological context that is required for a particular modification to occur. For example the KinasePhos (Huang *et al.* 2005a) tool predicts kinase-specific phosphorylation sites. Finally prediction tools are of course also useful in annotating potential modification sites for organisms that are scientifically interesting but unlikely to attract the funding required for large-scale proteomic analysis.

In addition to traditional prediction algorithms (which are briefly discussed below) there is also the notion of cross-annotation (Farriol-Mathis *et al.* 2004). Cross-annotation is the process of transferring feature annotations between proteins. Swiss-Prot curators carry out this process when new entries are added to the database (Farriol-Mathis *et al.* 2004). Of those PTM annotations imported into the PTMDB from Swiss-Prot, 23% have the evidence qualifier “By similarity” (which indicates a cross-annotation). Figure 26 shows a brief comparison between traditional prediction software and the process of cross-annotation.

To predict a protein feature *de novo* – requires the construction of rules that determine when a feature annotation applies and when it does not. Machine learning algorithms are now frequently being used, which are able to ascertain the rules that determine the designation of a particular residue as an acceptor or non-acceptor of a particular modification type.

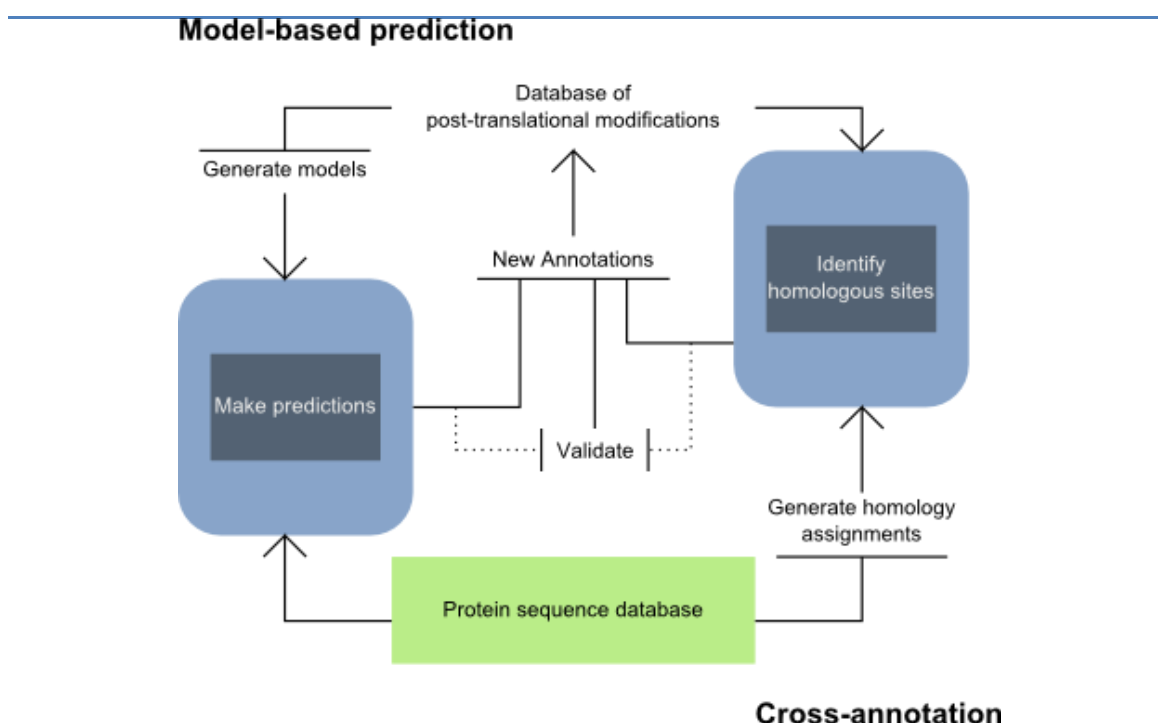


Figure 26: Comparison of model-based prediction with cross-annotation.

Section 4.2.2(a) Examples of prediction programs

The most widely used machine learning algorithms that have been used to date include: Support Vector Machines (SVMs) (EnsemblGly (Caragea *et al.* 2007), Profile Hidden Markov Models (KinasePhos (Huang *et al.* 2005b) and Neural networks (NetPhosK (Blom *et al.* 2004), (Senger and Karim 2008). In addition both rule based and PSSMs (Position Specific Scoring Matrices) have been utilised.

The Prosite database contains two different types of protein feature model (See Sigrist *et al.* 2010 and Sigrist *et al.* 2002a). This database includes models for predicting many different protein features including: binding motifs, modification sites, catalytic sites, etc. The two model types found in Prosite are termed patterns and profiles. Patterns are akin to regular expressions, which are frequently used in text mining; see Figure 27 for an example. They are composed of individual rules, which state what residues can and cannot occur at specific positions in a sequence motif. Patterns are constructed from multiple sequence alignments that aid in identifying conserved residues. Sequences either contain a match to a particular pattern or they don't; therefore they are qualitative. The prosite database is tightly coupled with the release of new

versions of the Swiss-Prot database. Swiss-Prot is scanned with the patterns present in Prosite to help identify those which occur often.

[RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y

Figure 27: Prosite, tyrosine kinase phosphorylation site pattern (PDOC00007). Pattern format is documented at the following URL <<http://expasy.org/prosite/prosuser.html>>; the tyrosine that is modified in this pattern is underlined and presented in a larger font.

Unlike patterns, that are specific to small highly conserved motifs, profiles are designed to match sequences across entire domains (which obviously includes regions of differing conservation). Like patterns they are constructed from an initial multiple sequence alignment (guided by 3D structures where these are available), which is converted into a profile hidden Markov model. Briefly these models are used to deduce the probability of a particular residue being present at a particular position in the domain; based on both a substitution matrix and the frequency of residues at each position in the original multiple sequence alignment. These individual probabilities are converted into a score for the whole aligned region – similar to a BLAST e-value; profiles are therefore quantitative.

ProRule (Sigrist *et al.* 2005) is a database of rules in UniRule format (<http://www.expasy.org/sprot/hamap/unirule_manual_short.html>) that are designed to aid in the annotation of genomes. The rules have been designed to overlay information from patterns and other biological resources onto profiles. Note that profiles do not include a facility to make the presence of certain residues mandatory for a domain to be matched against a query sequence; this limitation of profiles is addressed with ProRules. For example a ProRule can be created for a profile that represents a domain with a catalytic activity that makes the presence of residues key to this activity (derived from Prosite patterns and additional biological information) mandatory.

Profiles and ProRules are not directly related to the annotation of PTMs; however it is important to understand their use, as PTM sites can be made a mandatory requirement of a ProRule.

The dbPTM prediction tool chain (Lee *et al.* 2006) combines a number of different techniques to create models of PTM sequence motifs, which can be

seen as an evolution of the models found in Prosite. Like Prosite, the first stage is the construction of multiple sequence alignments for each motif. Next the tool utilises a technique called MDD (Maximal Dependence Decomposition), which clusters individual sequences that share the same dependencies between specific motif positions. Dependent positions can only be calculated when there is sufficient data to satisfy the statistical test that is used. The multiple sequence alignments from each MDD/raw cluster are then converted into a profile hidden Markov model. Profile hidden Markov models are more powerful than PSSMs as they are able to incorporate a model of how protein sequences evolve.

Section 4.2.2(b) *Positive and negative datasets*

The prediction of a protein feature usually starts with the acquisition of experimentally determined annotations for the feature of interest. When the protein feature is the presence of a specific PTM, this so called ‘positive dataset’ is composed of the known acceptor sites for this modification.

The Swiss-Prot database is commonly used to source the positive dataset used to train a PTM prediction model (Lee *et al.* 2006). In addition some databases, which are specific to a particular class of PTM, have also been used. For example the dbPTM prediction tool chain also includes annotations from the Phospho.ELM ((Diella *et al.* 2008a) - phosphorylation resource) and O-GLYCBASE ((Gupta *et al.* 1999) - O-linked glycosylation) databases (Lee *et al.* 2006).

Prediction methodologies also require a set of negative annotations – which are used in conjunction with the positive dataset to train the model on. Such datasets are much harder to generate for the prediction of PTMs owing to a lack of experimentally determined non-acceptor sites (Farriol-Mathis *et al.* 2004).

The Swiss-Prot database allows for the annotation of a non-acceptor site. The incorporation of such sites from the Swiss-Prot database into the PTMDB has already been discussed. In the PTMDB, there are only 135 such annotations compared with the 186,408 acceptors sites imported from the Swiss-Prot

database. Clearly there are not enough experimentally determined non-acceptor sites to form a negative dataset.

Previously published PTM prediction tools have been forced to attempt computationally to define non-acceptor sites. Farriol-Mathis *et al.*, 2004 warned the community that a predicted non-acceptor site must be “plausible” – it is no good including non-acceptor sites that would never be within the same compartment as the required enzyme(s) for example.

The negative dataset constructed by Lee *et al.* 2006 only utilises predicted non-acceptor sites from proteins that have experimentally determined acceptor sites at different positions; they created one negative dataset for each PTM class they were interested in. By only including such non-acceptor sites in this context, they guaranteed that the sites would be available to the necessary enzymes. Lee *et al.* 2006 state that they are making the assumption that all acceptor sites on such proteins have been characterised. They are also making the implicit assumption that the sites are surface accessible to the necessary enzymes at some point during their lifetime.

The authors of the PHOSIDA phosphorylation prediction SVM constructed a much simpler negative dataset. This dataset was composed of randomly chosen sites which were not in their positive acceptor site list (Gnad *et al.* 2007). Obviously this method doesn’t take into account the plausibility of the sites they have chosen.

Section 4.3 *Cross-annotation protocol*

The first step in cross-annotating modification annotations between homologous proteins is to create a pairwise sequence alignment. To increase the likelihood of cross-annotated sites being correct, Swiss-Prot curators have decided only to transfer annotations between proteins from closely related species (Farriol-Mathis *et al.* 2004). This policy could have been directly applied to this procedure using the orthology annotations incorporated into the PTMDB. However as orthology assignments have only been made between 30 species, this would rather limit the scope of the cross-annotation process. On the

assumption that PTMs are more likely to be conserved in conserved domains, it was decided to transfer annotations based on domain annotations instead.

Section 4.3.1(a) *Domain constraints*

The Pfam project (Finn *et al.* 2008) contains protein domain annotations for UniProtKB, which are produced by the HMMER tool (<http://hmmer.janelia.org/>). Pfam is composed of two datasets. PfamA is the gold standard protein domain dataset derived from manually aligned seed sequences. PfamB supplements the PfamA database providing automatically detected domains not present in PfamA. The PfamA domain annotations have been incorporated into the PTMDB as well as the provided domain alignments provided by this project. These domain alignments are used to identify homologous residues between proteins during the cross-annotation process. The PfamB dataset was not incorporated into the PTMDB for two reasons: i) it was undergoing extensive revision at the time of incorporation and ii) the domain annotations are, by definition, less reliable than those in PfamA.

Section 4.3.1(b) *Positive dataset*

The positive dataset represents all those modification annotations that are going to be available for cross-annotation to homologous sites. PTMs from the PTMDB, which had been extracted from the Swiss-Prot or Phospho.ELM databases, were used as the positive dataset. In the future support will be added for the PDB2Linucs dataset. It was decided only to cross-annotate sites that were annotated on the protein sequence listed for each UniProtKB entry. Therefore all those modifications annotated on splice variants were removed from this positive dataset.

Section 4.3.1(c) *Query dataset*

The use of pre-computed sequence alignments from the PfamA project significantly reduces the computational requirements of this cross-annotation process. However some aspects of the cross-annotation process are still quite computationally intensive. Performing a cross-annotation of all sequences in the PTMDB was determined to be an unattainable goal given the available hardware. It was also considered unnecessary to compare all 6,074,524 sequences in the PTMDB – as it has already been shown that a great deal of

redundancy exists. It was therefore decided that TrEMBL entries in the PTMDB proteome sets, and all those Swiss-Prot entries in the PTMDB, would be included in the positive set. In total the dataset contained 2,768,148 UniProtKB entries.

Section 4.3.1(d) *Cross annotation procedure*

PTM sites are first overlaid onto the PfamA alignments, which have been imported into the PTMDB. The alignment column, to which a site maps, is used to identify homologous sites in other proteins. Modifications that align to the same PfamA domain are analysed together. For each domain all proteins in the positive set are selected that are annotated with this domain. For every combination of annotation and protein, the algorithm checks if the target annotation is compatible with the residue found in the query sequence. A compatible site is identified by checking in the PTMDB vocabulary whether the given modification can occur on the query residue. It seemed logical that modified residues may be constrained during evolution only to be mutable to other residues that are compatible with the same modification. As an initial test of this idea the algorithm was set up so that all phosphorylated residue annotations could be interchanged. For example a phosphothreonine annotation mapped to a serine residue on a query would produce a phosphoserine annotation.

The cross-annotated Swiss-Prot annotations are considered to be somewhat reliable because they are manually checked for their suitability for the new protein. Without such manual checking an automatic method must have some way of scoring cross-annotations. An obvious method would be to run prediction programs on the same sequences and compare the cross-annotations to the predictions. Although this would be the best method it would be both: a) limited to modification classes for which prediction tools exist (and are not biased to particular organisms), and b) would be extremely computationally intensive. The cross-annotation algorithm has been designed to generate sequence windows around cross-annotated sites – between the target and query sequence. Sequence windows are created that extend left and right from the modified residue by 1, 2, 4 and 6 residues. In these

sequence windows a PSI is calculated and recorded for the match. Where a modification is closer than the extension length to the C/N terminal of the protein the window is extended by the missing amount in the opposite direction – this matches the technique used in dbPTM (not the PTMDB being discussed here).

The cross-annotation procedure was implemented on a relatively small HPC – with 9 nodes each equipped with two AMD 64bit Opteron processors and 2GB RAM. The procedure took 7½hrs to cross-annotate proteins in the Swiss-Prot database and 17hrs for those in TrEMBL.

Section 4.4 *Results*

Section 4.4.1 Incorporation of Pfam

Pfam annotations were only imported for Swiss-Prot entries and those TrEMBL entries included in the proteome sets, which have been incorporated into the PTMDB. Table 28 displays the number of proteins that had at least one Pfam domain annotation as well as the total number of annotations imported. Import statistics are shown in tables a) and b) in Table 28 for all UniProtKB entries referred to in the PTMDB proteome sets. Tables c) and d) show the same statistics for all Swiss-Prot entries, and table e) shows totals for these four tables.

The proteomes in the PTMDB are composed of 297,340 Swiss-Prot entries and 2,470,808 TrEMBL. 91% of the Swiss-Prot entries have at least one Pfam domain annotation compared with 62% of those from TrEMBL. The Pfam database used was not based on the same version of UniProtKB that was used to populate the PTMDB. It was therefore necessary to check that the UniProtKB entry sequence versions did not differ. A change to the protein sequence makes it impossible to know in which alignment column a modified residue is present. The Pfam annotations for 2,931 Swiss-Prot and 6,041 TrEMBL entries from the proteome were discarded because of a sequence version mismatch.

457,178 domain annotations have been imported for Swiss-Prot entries and 2,362,164 for TrEMBL entries. 71% of the proteins that have imported Pfam domains are from bacterial species. An additional 60,662 Swiss-Prot entries, not part of the proteome sets, have imported Pfam annotations.

In total, 1,878,940 proteins have had Pfam annotations imported and thus have the potential for PTM cross-annotations to be made onto them. Note that only ~0.5% of proteins and Pfam annotations have been lost because of sequence version mismatches.

Section 4.4.2 Target PTM set

The PTMDB contains 190,231 unique post-translational events. 52% (98,884) of these events have been overlaid onto Pfam domains. 7,834 Pfam domains have at least one acceptor site in the PTMDB. Table 29 displays the percentage of modified proteins in the PTMDB that have at least one of their annotations overlaid onto a Pfam domain, subdivided by PTM class. By protein abundance, phosphorylation and glycosylation are the most common modifications present in the PTMDB. For example; 66% of phosphoproteins and 81% of glycoproteins from the PTMDB are present in the target cross-annotation set. Many of the modification classes, which have a higher percentage inclusion than these two modifications, have extremely low abundance in the PTMDB.

Table 30 displays the percentage of modification sites specific to each PTM class which are annotated in Pfam domains. 54% of phosphorylation and 69% of glycosylation sites have been annotated to Pfam domains. Both tables show a similar distribution of numbers, most likely because most proteins only have one example of each modification.

Some modifications are specific to the C and N termini of protein sequences. These regions of a protein are less likely to be found in a PfamA domain – as shown in Figure 28.

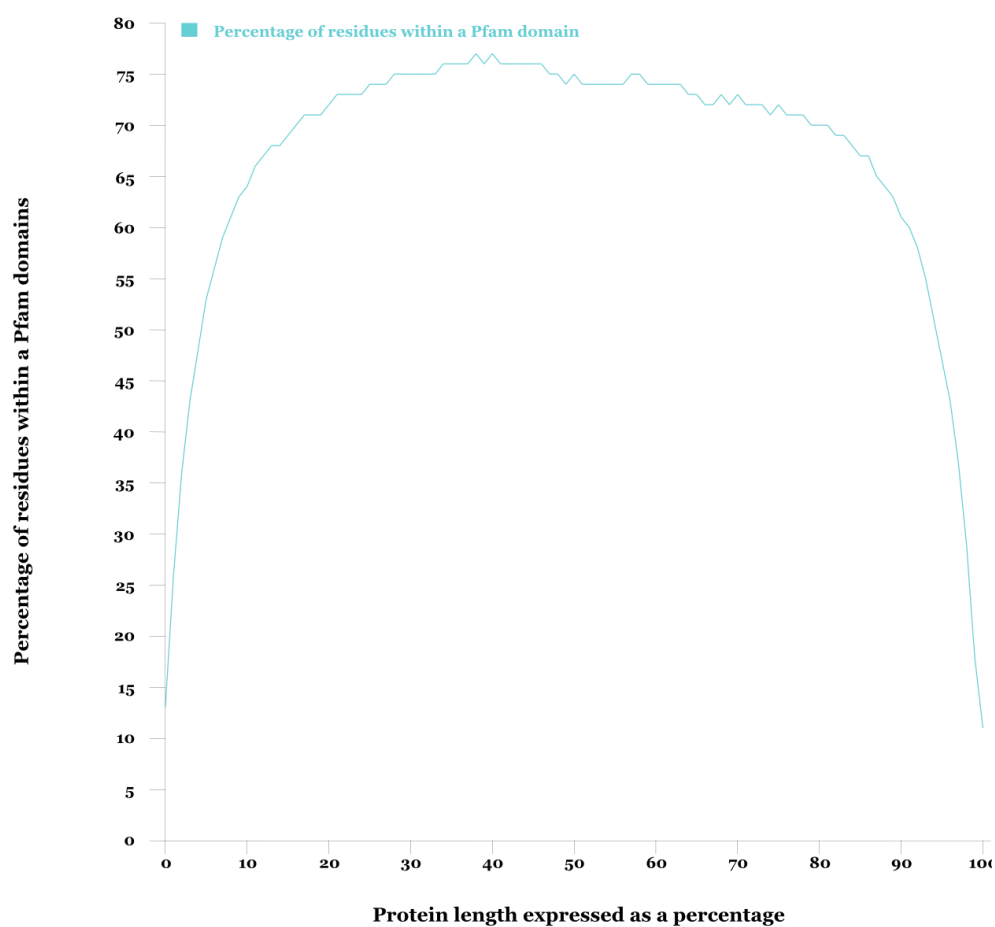


Figure 28: Limited conservation of N and C termini of proteins. Each position in an individual protein sequence has been converted into the percentage distance it is in the protein. The percentage of percentage-corrected positions found in Pfam domains – has then been plotted above.

Only 7% of prenylated proteins in the PTMDB are found in the target dataset. Figure 29 shows that prenylation sites are predominantly found in the c-terminus of proteins and that a very low fraction of these c-terminal sites are annotated on Pfam domains.

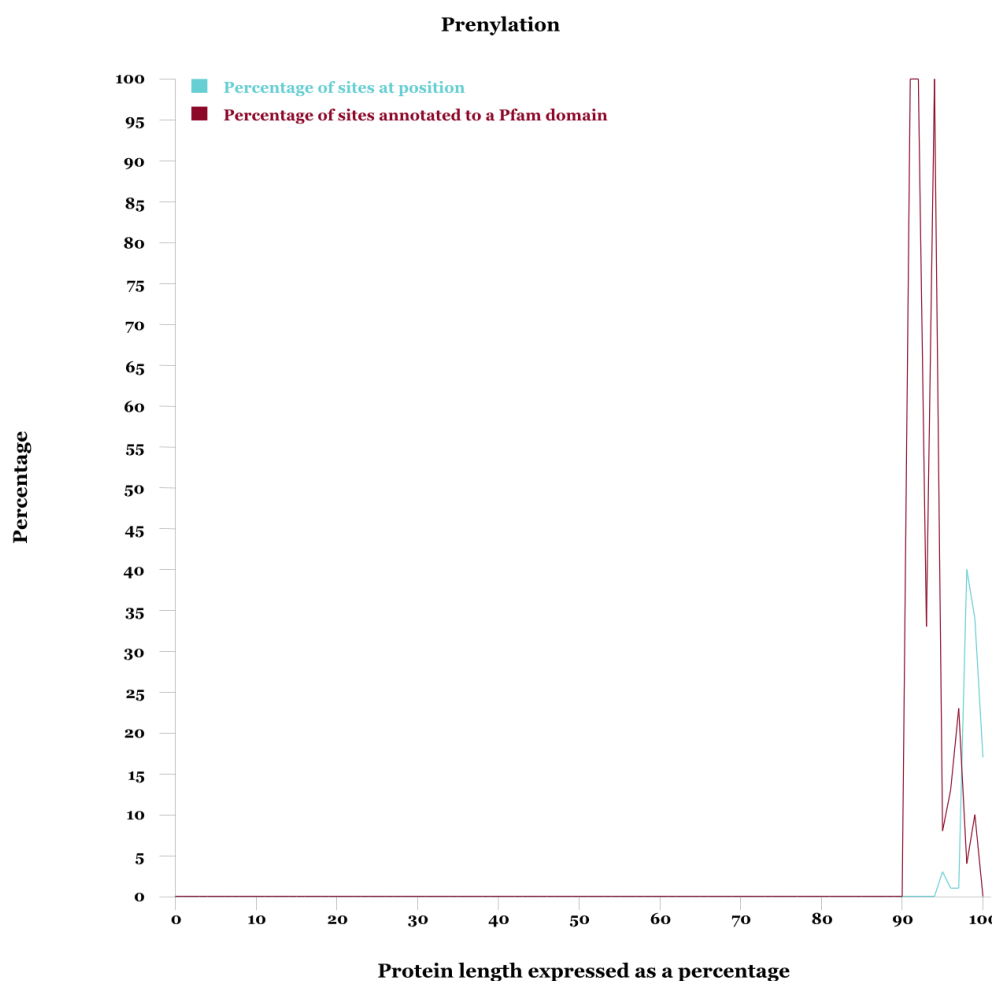


Figure 29: Distribution and annotation of Prenylation acceptor sites in the Pfam database.

In comparison Figure 30 shows that the GPI-anchor acceptor sites are similarly C-terminally distributed, although 41% of proteins with GPI-anchor modifications are included in the target set. Most modifications have a sequence distribution profile similar to that shown in Figure 30 for phosphorylation acceptor sites, which are evenly distributed throughout protein sequences.

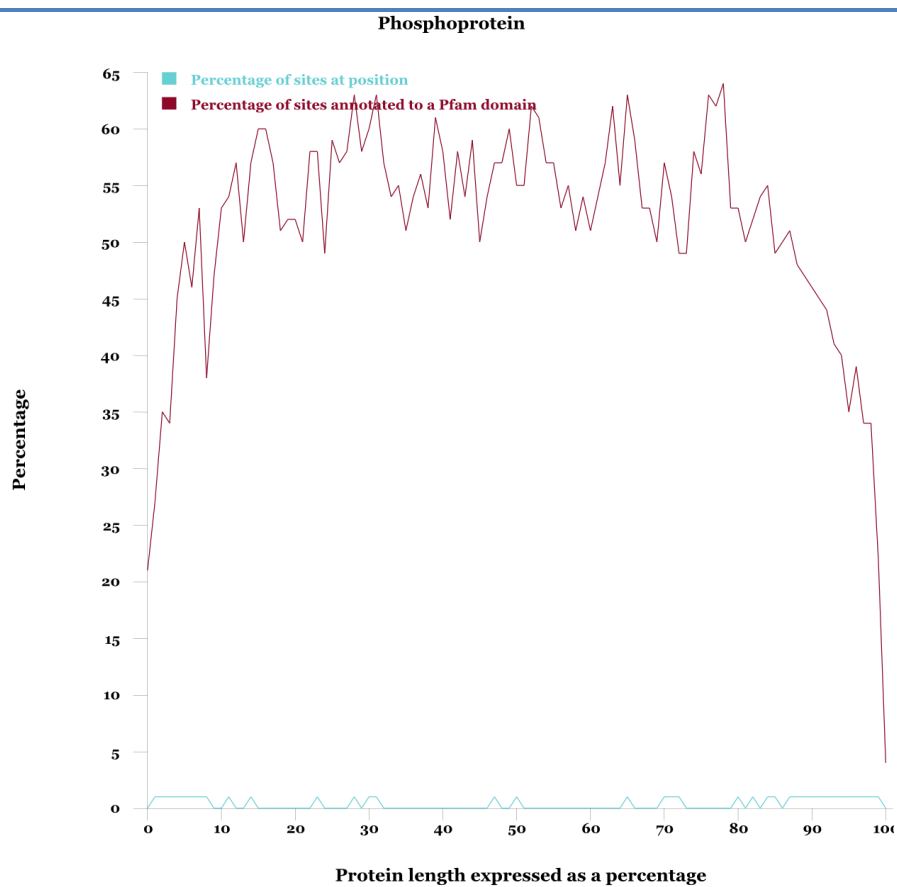
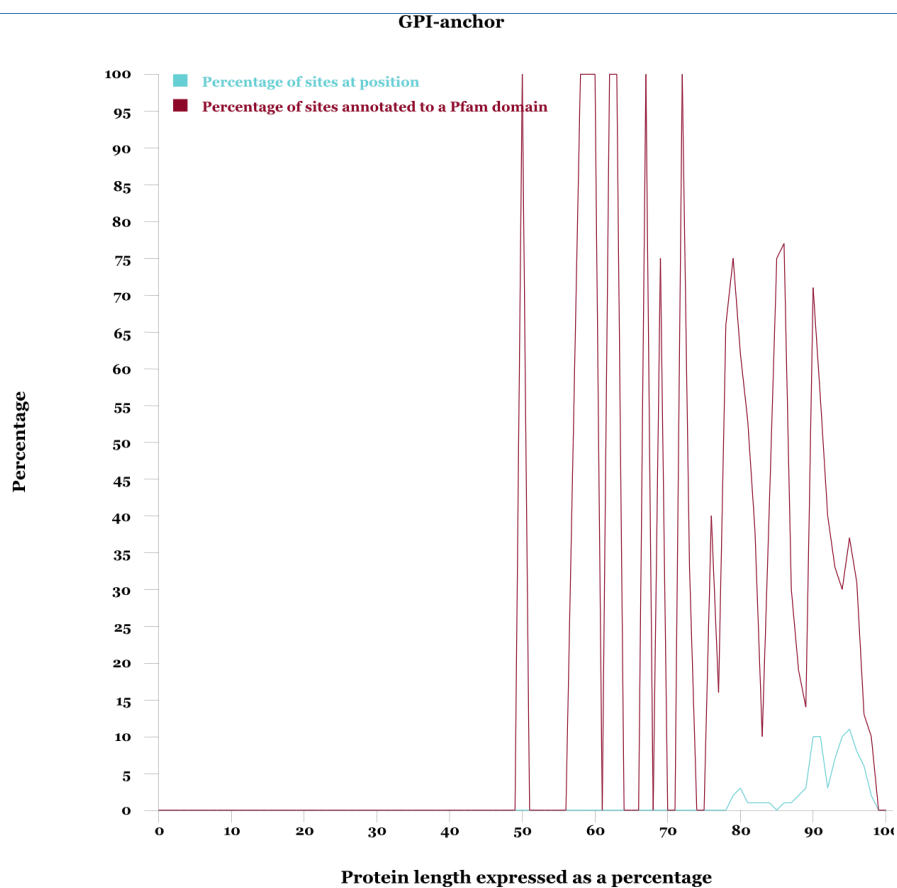


Figure 30: Comparison between the likelihoods of a residue at each position in a protein that is either (A) GPI-anchored or (B) phosphorylated, being annotated in a Pfam domain.

(a) Number of Pfam annotations imported and excluded for the query/target sets

Superkingdom	Kept		Excluded		Percentage removed	
	Swiss-Prot	TrEMBL	Swiss-Prot	TrEMBL	Swiss-Prot	TrEMBL
Eukaryota	154834	640427	3744	7385	2.36%	1.14%
Bacteria	280440	1626428	2073	2265	0.73%	0.14%
Archaea	17431	86552	114	74	0.65%	0.09%
None	4473	8757	21	18	0.47%	0.21%
Total	457178	2362164	5952	9742	1.29%	0.41%

(b) Number of proteins excluded

Superkingdom	Kept		Excluded		Percentage removed	
	Swiss-Prot	TrEMBL	Swiss-Prot	TrEMBL	Swiss-Prot	TrEMBL
Eukaryota	75290	353521	1628	4490	2.12%	1.25%
Bacteria	182161	1115798	1208	1508	0.66%	0.13%
Archaea	12324	60817	80	31	0.64%	0.05%
None	1753	7392	15	12	0.85%	0.16%
Total	271528	1537528	2931	6041	1.07%	0.39%

(c) Number of annotations excluded (Swiss-Prot)

Superkingdom	Kept	Excluded	Percentage removed
Eukaryotes	221898	4034	1.79%
Bacteria	297803	2107	0.70%
Archaea	18130	122	0.67%
None	16109	64	0.40%
Total	553940	6327	1.13%

(d) Number of Proteins (Swiss-Prot)

Superkingdom	Kept	Excluded	Percentage removed
Eukaryotes	117053	1812	1.52%
Bacteria	192871	1234	0.64%
Archaea	12790	88	0.68%
None	9476	47	0.49%
Total	332190	3181	0.95%

(e) Cross annotation (Total)

	Removed	Total	Percentage removed
Proteins	9222	1878940	0.49%
Annotations	16069	2932173	0.55%

Table 28: Comparison between the likelihoods of a residue, at each position in a protein, being either (A) GPI-anchored or (B) phosphorylated, being annotated in a Pfam domain.

PTM Class	% of proteins with at least one known acceptor site in a Pfam domain
Iodination	100.00%
Selenium	100.00%
ADP-ribosylation	100.00%
Organic radical	100.00%
Covalent protein-DNA linkage	100.00%
Pyruvate	99.08%
FAD	98.56%
Flavoprotein	96.10%
TPQ	95.24%
Oxidation	93.75%
Other	91.49%
FMN	90.91%
Glycosylation_C_Linked	90.63%
S-nitrosylation	88.24%
Glutathionylation	88.24%
Gamma-carboxyglutamic acid	87.50%
Glycosylation_N_Linked	81.27%
Glycosylation	81.03%
Peptidoglycan-anchor	79.70%
Nitration	75.51%
Hydroxylation	70.00%
Formylation	65.93%
Phosphoprotein	65.72%
Glycosylation_O_Linked	63.16%
Glycosylation_S_Linked	62.50%
Bromination	61.90%
Amidation	48.42%
Sulfation	47.02%
Palmitate	45.43%
Methylation	42.01%
GPI-anchor	41.67%
Glycoprotein	41.61%
Lipoprotein	40.23%
Nucleotide-binding	38.95%
Myristate	38.92%
D-amino acid	37.93%
Pyrrolidone carboxylic acid	37.86%
Hypusine	33.33%
Acetylation	30.88%
Citrullination	16.18%
Prenylation	9.45%
Covalent protein-RNA linkage	5.97%

Table 29: The percentage of proteins that have at least one of their known acceptor sites, for the given PTM class, in a Pfam domain.

PTM Class	Percentage of acceptor sites in a Pfam domain
Selenium	100.00%
Organic radical	100.00%
Covalent protein-DNA linkage	100.00%
Pyruvate	99.08%
FAD	98.56%
Flavoprotein	96.14%
Oxidation	94.59%
TPQ	93.02%
Gamma-carboxyglutamic acid	92.12%
FMN	91.18%
Other	89.47%
S-nitrosylation	85.00%
Peptidoglycan-anchor	79.70%
Nitration	77.19%
ADP-ribosylation	77.07%
Glycosylation_C_Linked	74.04%
Glutathionylation	71.43%
Glycosylation_N_Linked	69.69%
Glycosylation	69.09%
Formylation	65.47%
Glycosylation_S_Linked	62.50%
Hydroxylation	60.97%
Bromination	54.17%
Phosphoprotein	53.65%
Amidation	52.88%
Glycosylation_O_Linked	52.47%
Palmitate	48.61%
Iodination	47.06%
D-amino acid	44.30%
Sulfation	42.21%
GPI-anchor	41.67%
Glycoprotein	41.61%
Pyrrolidone carboxylic acid	40.84%
Lipoprotein	40.39%
Nucleotide-binding	38.95%
Myristate	38.94%
Methylation	37.51%
Hypusine	33.33%
Acetylation	32.07%
Citrullination	30.40%
Prenylation	7.15%
Covalent protein-RNA linkage	5.63%

Table 30: Percentage of acceptor sites in a Pfam domain.

Section 4.4.3 Cross-annotation results

Cross-annotations have been made between proteins regardless of how taxonomically close their parent species are. Cross-annotations have been made between, for instance, bacterial and eukaryotic species. Initially only

cross-annotations that have been made between species of the same super-kingdom will be described. A brief overview of the cross-annotations made between species of different super-kingdoms – including those taxa not annotated to any super-kingdom (e.g. viruses) – will then follow.

Section 4.4.3(a) *Window analysis*

As previously stated the cross-annotation protocol calculates the sequence identity in sequence windows of various lengths centred on sites of cross-annotation. These have been calculated to provide two parameters to which thresholds can be used to constrain the cross-annotation set. The annotation protocol also calculates sequence windows around homologous sites that already contained identical annotations before this cross-annotation procedure was carried out. Figure 31 plots the percentage of these sites that are lost as the PSI threshold is increased. This figure also shows the effect of varying the sequence window length.

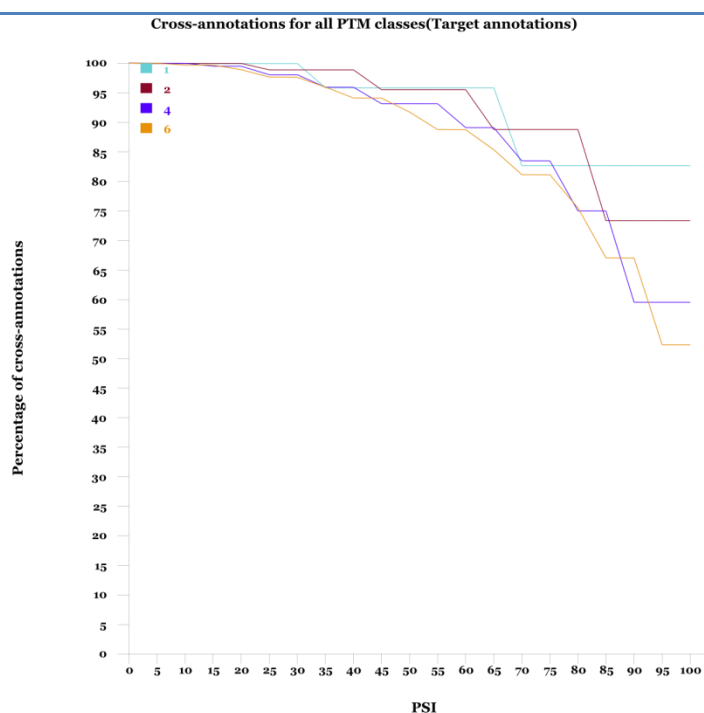


Figure 31: Conservation of the regions surrounding homologous PTM sites in the target dataset. Each line represents a different extension length to the left and right of the PTM site; the number of amino acids compared is twice the extension length. The y-axis represents the percentage of target PTM sites with at least one homologous partner in the target dataset; for each such site the highest observed PSI between it and all homologues is recorded for each extension length. The x-axis represents a \geq PSI threshold which is used to show what percentage of PTM sites in the target dataset with at least one homologue had a recorded PSI \geq a given value. For example ~50% of target sites (that had at least one homologous partner) had a recorded PSI of 100 in a sequence window of 12 amino acids.

Such a plot provides only a rough guide as to the sequence similarity which is observed around currently annotated PTMs in the PTMDB. Plots were also produced for specific PTM classes. Those for glycosylation and phosphorylation looked very similar to that in Figure 31. This plot shows that a large percentage of sites are still present even with quite a high PSI threshold – this suggests that the dataset being compared contains a high proportion of closely related sequences. As you might expect; as the PSI threshold is increased, the number of sequences lost increases quite dramatically.

This plot can be compared with that in Figure 32 which shows the same information for all those cross-annotations made for proteins in the Swiss-Prot database (note that the plot for TrEMBL was extremely similar).

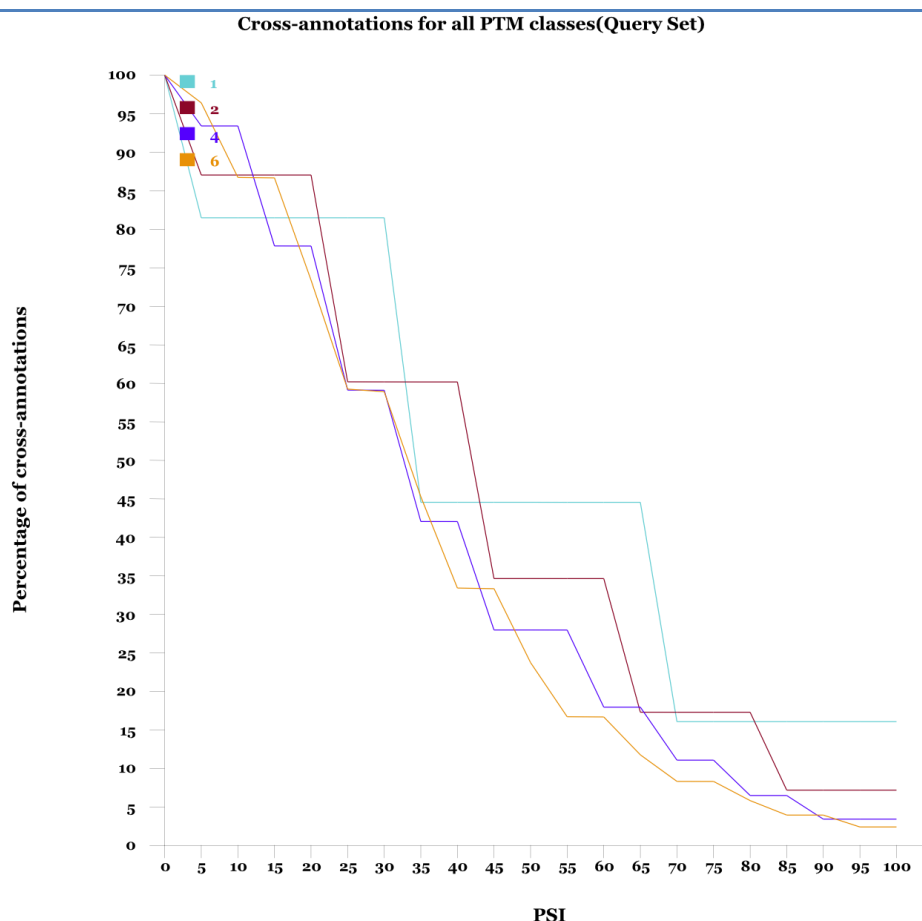


Figure 32: Sequence window analysis of the Swiss-Prot cross-annotation set.

This plot shows a much steeper loss of cross-annotations as the PSI is increased. As would be expected the cross-annotation protocol appears to be comparing sequences with a much greater degree of diversity compared to the

target set. This emphasises the need to apply a threshold to both the sequence window length and PSI.

Section 4.4.3(b) *Intra-super-kingdom predictions*

Table 31 shows the total number of cross-annotations that have been made. In total 54,453 Swiss-Prot and 262,265 TrEMBL entries now have PTM annotations that did not before. Between both databases 2,301,107 individual annotations have been added. These numbers drop significantly when a 70 PSI cut-off is applied to the new annotations with a sequence length of 12 aa (extension length 6). After applying these thresholds, 15,961 Swiss-Prot and 62,450 TrEMBL entries still have cross-annotations that did not have any modification annotations before this process was carried out. The total number of individual modification sites dropped to 163,763. Unless otherwise stated, all remaining statistics have been used after a 70 PSI cut-off was applied to a sequence window of 12aa.

	Number of Proteins		Number of modifications	Number of acceptor sites
	On existing	New		
Complete				
Swiss-Prot	20,340	54,453	620,042	576,632
TrEMBL	24	262,265	1,139,179	1,113,662
Total	20,364	316,718	2,301,107	1,690,294
70% SI 12aa				
Swiss-Prot	7,811	15,961	60,610	57,400
TrEMBL (70 6)	8	62,450	103,153	100,827
Total	7,819	78,411	163,763	158,227

Table 31: Gross number of cross-annotations made in the Swiss-Prot and TrEMBL database. The top three rows represent the whole cross-annotation dataset before any constraints are applied. The final three rows represent those annotations with > 70 PSI in a window of 12 aa.

A breakdown of the number of proteins that received cross-annotations for each PTM class is shown in Table 32. The PTM classes that were most abundant in the target annotation set are also the most abundant in the predicted set. Phosphorylation and Glycosylation were the most abundant cross-annotations when ranked according to the number of proteins with a new annotation. Of the 78,411 proteins that received a cross-annotation, 55,466 (70%) had at least one phosphorylation annotation and 16,407 (20%) received a glycosylation annotation.

PTM Class	Swiss-Prot		TrEMBL	
	On Existing	New	On Existing	New
Phosphoprotein	2742	12795	7	39922
Glycosylation	2987	1863	0	11557
Glycosylation_N_Linked	2785	1818	0	11335
Methylation	185	2947	0	4046
ADP-ribosylation	0	274	0	3213
Acetylation	231	1512	0	1823
Other	0	489	0	1212
Lipoprotein	2	269	0	957
Palmitate	2	109	0	844
Flavoprotein	0	0	0	753
FAD	0	0	0	587
Nitration	4	849	0	541
Hydroxylation	57	875	0	465
Nucleotide-binding	0	0	0	442
Organic radical	0	0	0	370
Pyruvate	0	2	0	339
Oxidation	0	33	0	311
S-nitrosylation	0	709	0	274
Glycosylation_O_Linked	32	302	0	269

Table 32: Number of proteins with cross-annotations for each PTM class.

A breakdown is now provided of the cross-annotations made for species in the three different super-kingdoms.

Bacteria

Before cross-annotation, 799 species/strains had at least one PTM annotation in the PTMDB. This number increased slightly to 884 after this process. Table 33 shows the number of proteins that have been annotated with each PTM class for all bacterial species in the query protein set. The most abundant modification cross-annotation was for phosphorylation with 20,236 proteins being annotated with at least one new acceptor site (note that 40% of the original bacterial annotations were for this modification). The number of phosphorylated proteins for bacteria has increased almost 9-fold. A significant increase in the number of methylation and ADP-ribosylation sites was also observed.

PTM Class	Swiss-Prot		TrEMBL
	On Existing	New	New
Phosphoprotein	122	1844	18270
Methylation	0	1612	3085
ADP-ribosylation	0	1	2836
Other	0	270	920
Palmitate	0	25	612
Lipoprotein	0	25	612
Flavoprotein	0	0	570
Nucleotide-binding	0	0	430
FAD	0	0	416
Organic radical	0	0	368
Oxidation	0	31	278
Acetylation	0	16	255
Pyruvate	0	2	217
FMN	0	0	154
Hydroxylation	0	4	137
Peptidoglycan-anchor	0	0	118
S-nitrosylation	0	3	78
Glutathionylation	0	3	78

Table 33: Bacteria cross-annotation statistics – some PTM classes have been excluded.

Archaea

Before the cross-annotation process the PTMDB contained 666 modifications to sites in archaeal species – 385 of which mapped to positions in PfamA domains. Of the three super-kingdoms; archaeal species have by far the fewest modifications in the PTMDB. As previously noted; 86,230 proteins received new annotations from the cross-annotation procedure – 196 of these were archaeal proteins.

Eukaryotes

Eukaryotes account for 95% of the annotations present in the target set of PTMs used for the cross-annotation process. This contained annotations for 2,047 species – the cross-annotation has resulted in an additional 713 species having at least one annotation. Table 34 shows that 35,203 eukaryote proteins have received new phosphorylation annotations and 16,374 new glycosylation annotations.

As already stated, phosphorylation and glycosylation are the most abundant cross-annotations. Figure 33 shows the distribution of these cross-annotations

between Eukaryote species. It is interesting to note that a few species that did not prominently feature in the PTMDB before this process now have a significant number of their proteins annotated with at least one annotation. For example, the PTMDB contained annotations for seven *T. nigroviridis* proteins; this cross-annotation process has increased this number to 3,702. Another fish, *D. rerio*, had no annotations in the target set; after cross-annotation, 2,853 proteins had at least one annotation. In addition this figure shows that a lot of cross-annotations have been made for *H. sapiens*, *M. musculus*, and *Xenopus laevis*.

The distribution of phosphorylation cross-annotations between the different types of phosphorylation between eukaryote species is shown in Figure 34.

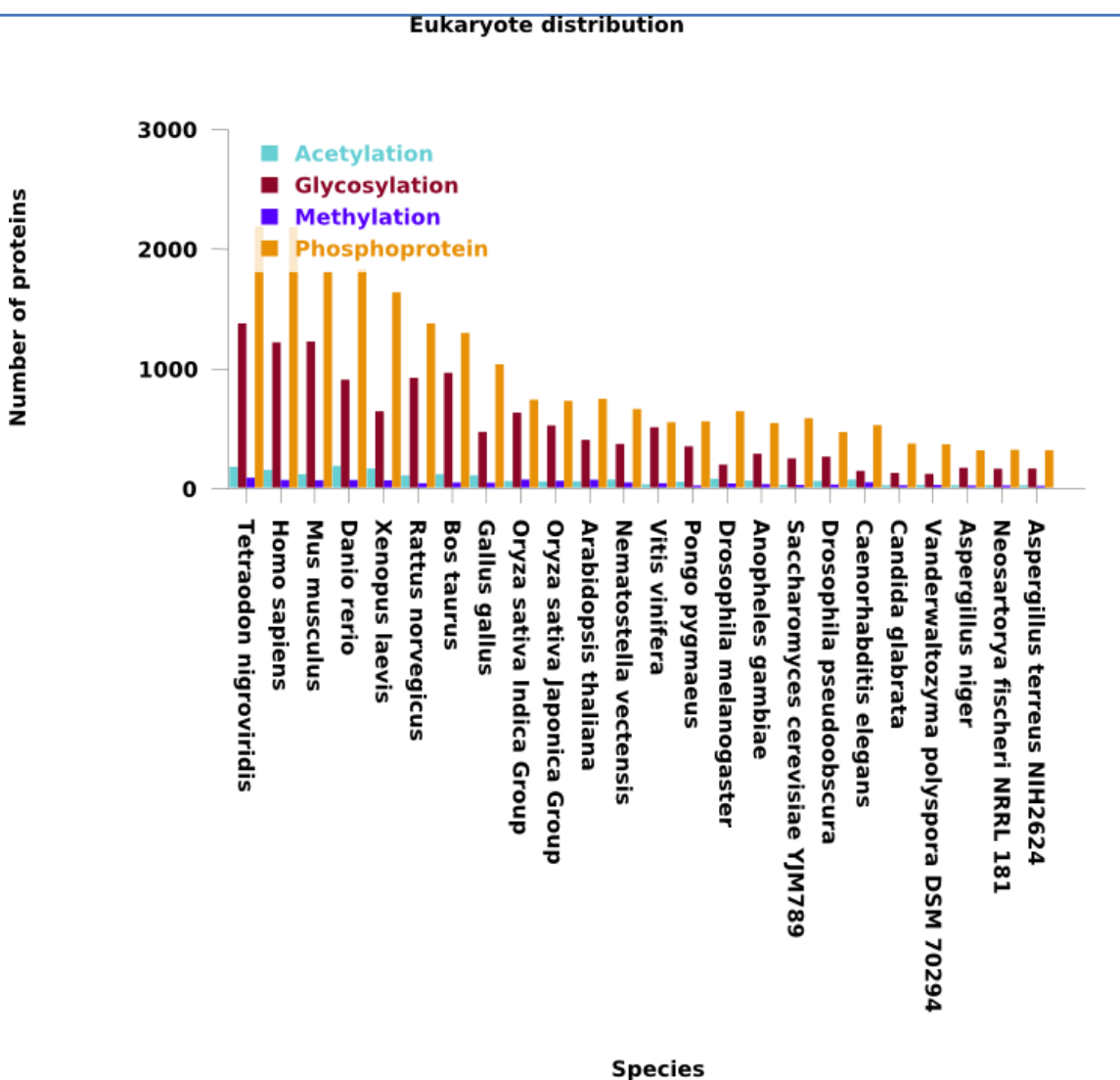


Figure 33: Distribution of predicted Glycosylated, N-linked Glycosylated and Phosphorylated proteins between different Eukaryote species. Note that all species with less than 500 proteins that have been cross-annotated have been excluded from this figure.

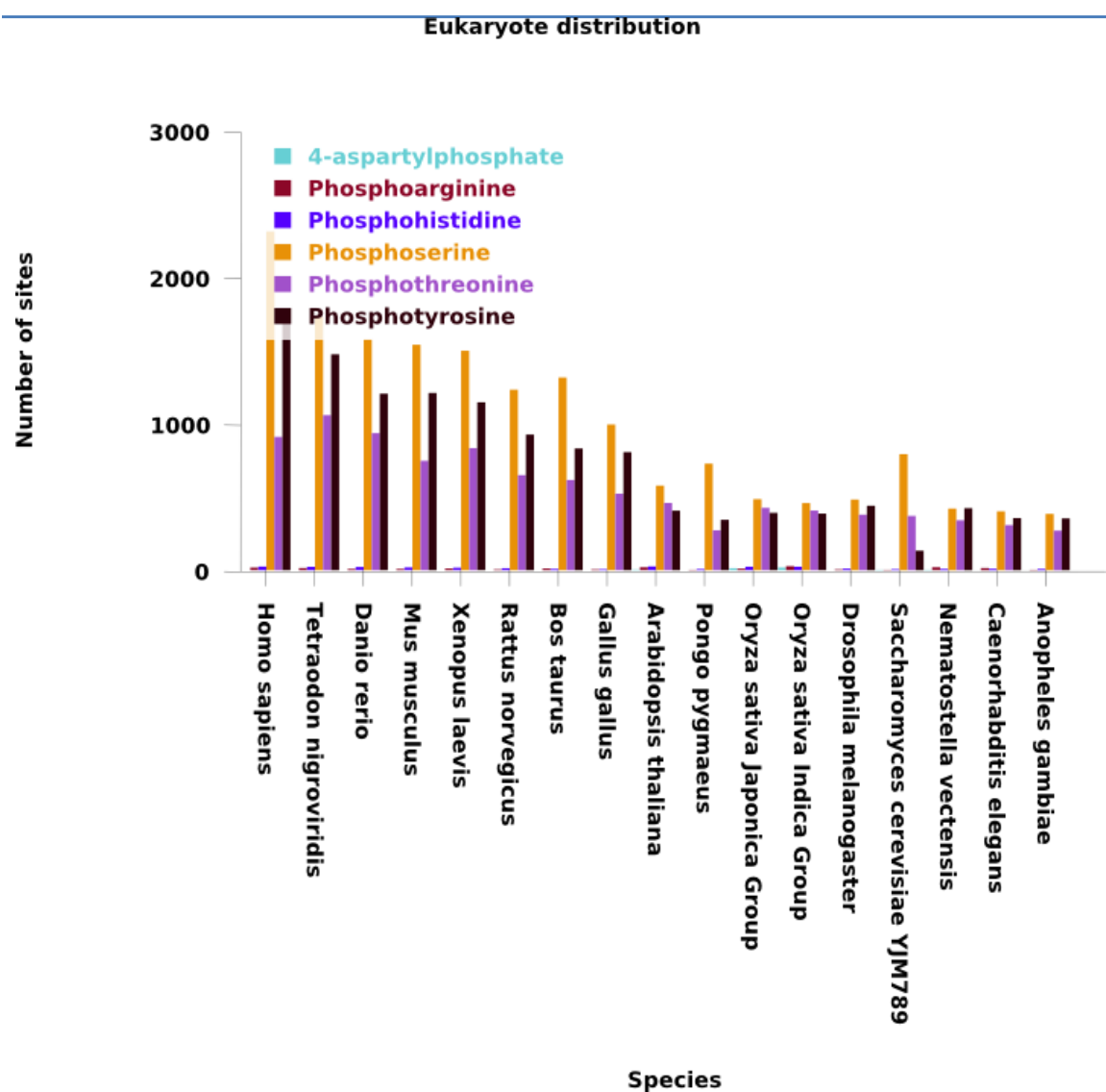


Figure 34: Distribution of phosphorylation cross-annotations between the different types of phosphorylation, grouped by species. The number of residues that have been modified is also displayed. Note that species with less than 1000 phosphorylation cross-annotations are not shown

PTM Class	Swiss-Prot		TrEMBL	
	On Existing	New	On Existing	New
Phosphoprotein	2620	10951	7	21625
Glycosylation	2987	1862	0	11525
Glycosylation_N_Linked	2785	1817	0	11307
Acetylation	231	1495	0	1551
Methylation	175	1335	0	940
Nitration	4	849	0	541
ADP-ribosylation	0	273	0	347
Lipoprotein	2	244	0	342
Hydroxylation	57	871	0	328
Glycosylation_O_Linked	32	302	0	265

Table 34: Eukaryote cross-annotation statistics. Note that for brevity not all keywords are shown in this table.

Section 4.4.3(c) *Inter-super-kingdom cross-annotations*

Some PTM classes are thought to be exclusive to particular lineages; for example glycosylation is almost exclusively associated with Eukaryotes. This is one reason why cross-annotating modifications between species that belong to different super-kingdoms might not be a valid procedure. However it should be pointed out that some modification sites have been identified that are conserved across super-kingdoms (Macek *et al.* 2008). In the target annotation set, 7% of the PfamA alignment positions annotated with a modification in Bacteria shared the same annotations with Eukaryote proteins.

In addition it is important to note that many Prokaryotes have now been recorded as being able to glycosylate some of their proteins (Hitchen and Dell 2006). It is important to stress again at this point that the PTMDB does not contain detailed glycan structures; therefore glycosylation cross-annotations do not imply anything about the conservation of glycan structures.

Although some conservation of modification sites between species belonging to different super-kingdoms has been experimentally confirmed; the majority of sites appear unlikely to be conserved. Therefore the inter-super-kingdom cross-annotations should be treated as a dataset that can be further interrogated and refined.

Table 35 summarise the inter-super-kingdom cross-annotations that were performed and the direction of the transfer. This table includes all inter-super-kingdom cross-annotations and not just those that could satisfy the thresholds applied to the intra-super-kingdom annotations – previously described. Whilst the results of this table are interesting they are also a warning. This table shows that without applying any thresholds an extremely large number of cross-annotations are selected for inter-super-kingdom matches. Note that this table shows the number of alignments that generated a cross-annotation and not the number of unique sites for which an intra-superkingdom cross-annotation was performed.

PTM Class	Donor->recipient					
	e->b	b->e	e->a	a->e	a->b	b->a
Phosphoprotein	205,924	12,758	12,519	137	11	1,320
Glycosylation_N_Linked	161,509	38	7,506	19	47	30
Acetylation	26,101	477	1,295	20	0	327
Nitration	4,906	0	275	0	0	0
Methylation	3,237	199	276	990	1,797	156
Hydroxylation	1,616	248	31	0	0	42
Glycosylation_O_Linked	1,280	14	23	0	0	0

Table 35: Inter-super-kingdom predictions. The number of acceptor sites that only have evidence from a species of a different super-kingdom are shown above. The direction of annotation is shown in the form donor->acceptor – where; a=Archaea, b=Bacteria and e=Eukaryote. For example the first cell contains the number 205,924, which indicates the number of phosphoprotein annotations that have been transferred from Eukaryotic species to bacterial ones.

Section 4.4.4 Comparison to UniProtKB release 2010_12

The target PTM dataset was derived from annotations released as part of Swiss-Prot release 55.3 (published in August 2008) and Phospho.ELM release 7 (published in July 2007). At publication the most recent release of the UniProtKB was the December 2010 version (release tagged as 2010_12). The 2010_12 version has been used to create an updated version of the PTMDB – which has been searched for the previously discussed cross-annotations.

The PTM Browser web services API was used to retrieve cross-annotation sets that were specific to Eukaryotes, Mammals, Bacteria and Archaea. The web service was asked to exclude annotations that didn't match any in the target set (all those annotations imported from Swiss-Prot and Phospho.ELM) for the same taxonomic group. For example the returned Mammalian dataset didn't contain any cross-annotations derived entirely from non-Mammalian proteins.

Table 36 provides a breakdown of the number of proteins present in each taxonomic-specific cross-annotation dataset (extracted as described above). In order to understand how many of the cross-annotations are present in the 2010_12 release of Swiss-Prot – it is important to remember that some Swiss-Prot entries may have been removed, and some TrEMBL entries added, to Swiss-Prot; this table records such events. Furthermore it's important to only compare annotations where the sequence version number is unchanged; again the same table records this information. Across all four datasets very few of the TrEMBL proteins with cross-annotations have been incorporated into the latest

release of the Swiss-Prot database. This appears to indicate that although the PTMDB is based on an older version of Swiss-Prot its cross-annotations to TrEMBL proteins may still represent a useful resource for scientists. Note that a relatively small number of the Swiss-Prot entries, that had cross-annotations, appear to have been removed from Swiss-Prot (196 for Eukaryotes). Some proteins have been excluded because of sequence version mismatches; for example 13% of TrEMBL and 5% of Swiss-Prot entries belonging to Eukaryotes were excluded. A very small number of cross-annotations had corresponding PTM sites (defined as a unique triplet of UniProtKB accession, residue number, and PTM class) in the 2010_12 Swiss-Prot release. For example of the 65,567 Eukaryote Swiss-Prot proteins there were only 1,891 matches; which is 0.028 matching sites per protein. It is not particularly surprising that so few sites match between the two datasets, especially when you consider that new annotations are added manually by curators.

Taxa	Database	No Proteins			
		Total	Present 2010_12 Swiss-Prot	Sequence version matches	Matching annotations
Eukaryotes	TrEMBL	189161	5590	4870	863
Eukaryotes	Swiss-Prot	68780	68584	65567	1891
Mammalia	TrEMBL	12248	920	720	228
Mammalia	Swiss-prot	32861	32753	30550	1403
Bacteria	TrEMBL	73168	1805	1680	393
Bacteria	Swiss-Prot	10513	10504	10456	12

Table 36: Number of Cross-annotations present in PTMDB that are also present in UniProtKB release 2010_12.

Section 4.5 Discussion

The primary aim of performing the cross-annotation procedure was to expand the PTMDB to include proteins that are in the TrEMBL database. A secondary objective was to identify and add any annotations that appeared to be missing in homologous sites between Swiss-Prot entries. These new annotations are of course not of the same high-quality associated with those in the Swiss-Prot (excluding those annotated as potential or probable) and Phospho.ELM databases. The assumption was made that the higher the local sequence identity between the target and query site, the more likely the query residue is to be modified. Therefore the PSI between the target and query sites in windows of various lengths was calculated for each cross-annotation. A PSI threshold

can therefore be used to compensate for the fact that these cross-annotations have not been manually curated like those in the two previously-mentioned databases. It is important to point out that when a cross-annotation takes place it is guaranteed to take place between homologous sites; assuming that the Pfam domain alignment is correct. Therefore there is no need for any type of alignment scoring metric to indicate if two regions are homologous.

It would be beneficial to know which sites match known recognition sites for corresponding enzymes. This could be accomplished by scanning the regions surrounding the cross-annotation sites with the recognition patterns from Prosite of such enzymes. Note that such patterns either match or don't match (e-values are only calculated for Prosite profiles); although Prosite does indicate if a pattern occurs frequently in the Swiss-Prot database.

Even without the aid of a prediction tool, such as Prosite, a high-quality dataset of cross-annotations can still be obtained by setting a high sequence identity threshold in a chosen window length. As an aid to selecting an appropriate threshold a comparison was provided between the sequence identities of homologous sites that already have the same modification annotations. This comparison demonstrated that in the PTMDB there was a relatively high degree of sequence identity between such sites. This could simply be an artefact of curators only annotating homologous sites where the sequence identity is high. After analysing this comparison, a somewhat arbitrary threshold of > 70 PSI in a sequence length of 12 amino acids was chosen to create a set of cross-annotations, which were discussed in the previous section – all results discussed below also used this threshold. 25% of cross-annotations were made between sequences with > 70 PSI in a window size of 12 aa.

The majority of new cross-annotations have been made for eukaryotic proteins. A much smaller number of bacterial and archaeal proteins have been annotated with PTMs. The large number of inter-super-kingdom cross-annotations (without thresholds applied) shows that residues which are modified (particularly those in eukaryotes) are homologous to those in proteins of different super-kingdoms. Such inter-super-kingdom cross-annotations have not yet been included in the PTMDB. They have been excluded on the basis that few

modification sites appeared conserved between proteins of different superkingdoms before this process was carried out, which suggests that what is being observed is a high degree of sequence conservation that doesn't necessarily correlate with the presence of a conserved modification. This conclusion isn't particularly surprising given that only modification sites in conserved protein domains have been cross-annotated. A rather shocking example of this is provided by the glycosylation annotations of eukaryote species that have been cross-annotated to bacteria. 479,161 bacterial residues received a glycosylation annotation; before this process was carried out, only 26 such residues had a glycosylation annotation. Note that although glycans have been widely observed on bacterial proteins (Hitchen and Dell 2006) they have not been reported to this level.

As a cross-annotation can only be made between residues annotated to the same domain, it is obvious that this procedure is vulnerable to missing domain annotations. A missing domain annotation would manifest itself as a modification not being conserved between two homologous positions. A related problem occurs when the domain annotations are based on a set of sequences that differ from those used in the PTMDB. For example, both Pfam and the PTMDB are based on sequences in UniProtKB. Issues appear when these two databases are based on different versions of UniProtKB. For example a modified protein in the PTMDB may lack domain annotations in the Pfam database, because they were simply missing from the version of the UniProtKB upon which it is based.

The Pfam database annotates domains on the primary sequence of UniProtKB entries. The UniProtKB and Pfam databases are not released at the same time it is therefore possible that the domain annotations were made on a sequence which is different from that in the PTMDB. Domain annotations which have a different sequence version number in the Pfam database compared with the PTMDB have not been incorporated into the PTMDB. Such annotation mismatches are however recorded, allowing future analysis software to take into account annotations that have been removed.

As already mentioned, Lee *et al.* 2006 created a negative dataset that was specific to each PTM class for which they made predictions. For each such class, they extracted residues from proteins with experimentally determined annotations of the given class that were themselves compatible with the same modification but had not yet been annotated as such. They are clearly implying that proteins that have known experimentally-determined PTM sites of a given class are likely to have had the majority of their PTMs characterised. Although the cross-annotation procedure did annotate some sites that would fit their definition of a non-acceptor site, these annotations are in the minority. In the future, groups might want to check the cross-annotation set with an appropriate selection of thresholds to identify sites that they could exclude from their negative dataset.

In conclusion, the cross-annotation procedure has resulted in an explosion in the number of annotations present in the PTMDB. These cross-annotation results can be constrained by sequence identity to provide a high quality dataset. 62,450 (70% SI threshold) TrEMBL entries now have at least one modification annotation in the PTMDB compared with just 204 previously (all of which were for phosphorylation, extracted from the Phospho.ELM database). In total 163,763 (70% SI threshold) cross-annotations have been incorporated into the PTMDB.

Chapter 5

PTM Browser

Section 5.1 *Summary*

There are currently few bioinformatics tools that allow users to interrogate post-translational modification datasets. In particular it is difficult for researchers outside the field of bioinformatics to ask simple questions regarding the conservation and co-occurrence of PTMs. This has necessitated the development of a novel tool and associated algorithms to answer such questions. The new tool described in this chapter – PTM Browser – is freely available for non-commercial use at the following URL: <<http://www.ptmbrowser.org>>. The PTMDB provides annotations to this tool, which includes the cross-annotations and CoPaO orthologue/parologue assignments discussed previously. Simple questions such as: what percentage of the *H. sapiens* phosphoproteome is conserved in *M. musculus* or which *H. sapiens* glycosylation sites do *M. musculus* lack can be asked. This tool should be extremely useful to scientists who find that the proteins of their organism of interest are predominantly in the TrEMBL database – as the PTMDB now contains annotations for a large number of such proteins. Finally a web service has been provided in the hope that bioinformaticians will be able to build on the analysis pipeline in PTM Browser.

Section 5.2 *Application implementation details*

PTM Browser has been written as a web application using the Ajax (Aynchronous JavaScript and XML) methodology. The client side of PTM Browser makes use of CSS (Cascading Style Sheets) and the JQuery (JQuery-Community 2010) JavaScript framework. The server side component of PTM Browser has been programmed on top of the MVC (Model View Controller) framework Ramaze (Fellinger 2010). This framework has been written in the Ruby language although PTM Browser has been deployed on the Java implementation of the Ruby language, JRuby (JRuby-Community 2010). This decision allows PTM Browser to access Java objects from its Ruby code. Additional graphing routines have been written in the Perl language utilising the Cairo graphics library (Cairo-Community 2010). As with the PTMDB already discussed, PTM Browser uses MySQL (Oracle 2010) (RDBMS) for data storage. PTM Browser is hosted on a Ruby web server that has Apache 2 (Apache-Community 2010) acting as its proxy to the outside world.

PTM Browser is divided into a client and server side component. The server side component exposes a public web service API (see Section 5.4) that can be used to create user interaction and data analysis software. The PTM Browser client is an example of an application that uses this API. The source code for PTM Browser has been made available at the following URL <<http://code.google.com/p/ptmbrowser>>. Further details on the implementation of PTM Browser are supplied as required throughout the rest of this chapter.

The PTM Browser client web interface is shown in Figure 35.

Section 5.2.1 PTMDB inclusion

The PTM Browser tool has been designed to utilise only a subset of the features available in the PTMDB that have thus far been described. Users are able to select from two datasets. The first, referred to as PTMDB_{Full}, contains all the annotations from the PTMDB with the exception of those derived from the PDB. The second, referred to as PTMDB_{PfamA}, contains only those annotations that have been successfully mapped onto a Pfam A domain. Note that this second dataset excludes all annotations imported from the PDB.

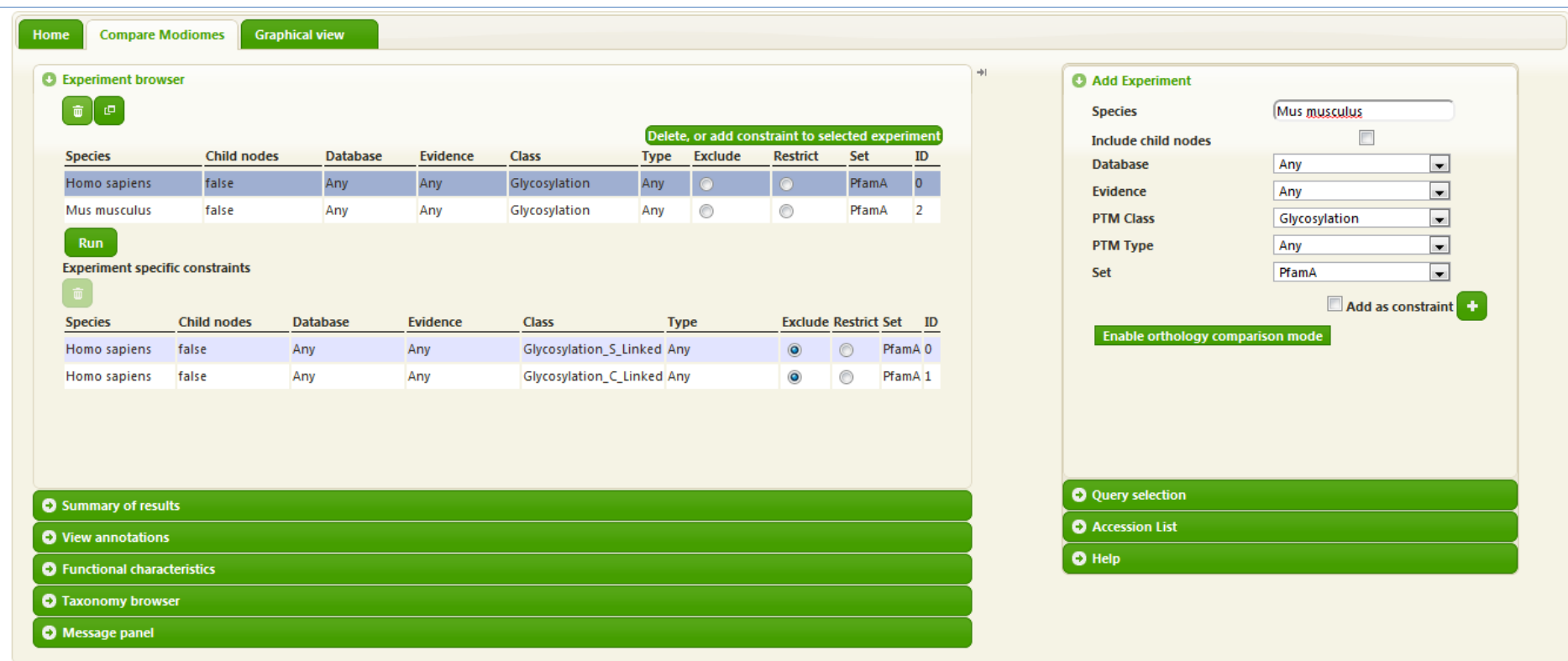


Figure 35: PTM Browser client web interface.

PTM Browser left accordion contains a number of expandable panels that show results and other information to users. The Experiment browser is shown, which displays the annotation sets that the user has selected. This component is also used to group, delete and apply constraints to experiments. The right accordion contains panels to add annotation sets to the query browser and select the query type

A comparison of the features provided by these two databases and the original PTMDB and the SwissProt database is shown in Table 37.

Note that permission has been obtained from the Phospho.ELM authors to distribute their data through PTM Browser. Additionally permission was required from the authors of the original InParanoid method to distribute the orthologue/parologue clusters produced by the CoPaO program. This was required as CoPaO makes use of the BLAST parsing routine (does not perform the actual detection of orthologues and in-paralogues) written by the authors of the InParanoid algorithm. Permission has been obtained to distribute the CoPaO assignments through PTM Browser.

Database	UniProtKB		Phospho.ELM	PDB-2-Linucs	Cross annotations		Pfam Annotations		Homology assignments
	SwissProt	TrEMBL			inter	intra	A	B	
SwissProt	✓	X	✓	X	X	✓	✓	✓	KEGG
PTMDB	✓	✓	✓	✓	✓	✓	✓	✓	CoPaO
PfamA	✓	✓	✓	X	✓	✓	✓	X	CoPaO
Full	✓	✓	✓	X	✓	✓	X	X	CoPaO

Table 37: Comparison showing the information present in Swiss-Prot, PTMDB and its subset databases: PfamA and Full, that are used by the PTM Browser application

During the development of PTM Browser it became evident that the PTMDB schema did not have the performance characteristics required to allow for certain queries to be performed in a reasonable time frame. A static representation of the PTMDB was therefore created, according to the schema shown in Table 38. At the time of writing PTM Browser was deployed on a virtual server with poor disk IO performance compared with that of a dedicated server. This issue was partly addressed by the new schema which converted many fields from a varchar type to an enum to reduce the size of the database – allowing the database to be loaded into a MySQL Hash table (stored in RAM). To reduce the number of table joins performed by PTM Browser, Pfam alignment coordinates are stored directly in this table. In addition the alignment window statistics, calculated for the cross-annotated PTM sites (discussed in Section 4.4.3(a), have been merged into the same table. Therefore this table will contain one row for each combination of cross-annotated PTM annotation and window alignment length.

Field name	Type
AnnotationId	int(11)
UniProtKBAccession	char(9)
Start	mediumint(8) unsigned
AlnPosition	smallint(5) unsigned
PfamAccession	varchar(8)
PtmDescription	enum('Phosphohistidine',...)
MethodDescription	enum('By similarity',...)
DatabaseName	enum('tr emblPfamA',...)
KwDescription	enum('Phosphoprotein',...)
NetId	int(11)
AlnLength	tinyint(4)
Identities	tinyint(4)

Table 38: PTM Browser PTMDB modified table structure.

Section 5.2.2 Pfam complications

It is stated that no two PfamA domain assignments may have overlapping alignment coordinates (Finn *et al.* 2010). However some PfamA assignments appear to violate this rule thus complicating any analysis that is dependent upon this rule. This central rule appears to be violated under two different conditions. To understand the first mechanism it is necessary to understand what the Pfam authors actually mean by an “alignment coordinate”. The HMMER package is used to produce possible domain annotations for protein sequences (see <<http://hmmer.janelia.org/>>and (Eddy 1998). HMMER outputs two aligned regions for each possible domain match: a low and high confidence region referred to as the “envelope” and “alignment” windows (Finn *et al.* 2010). An overlap is explicitly allowed between the envelope regions of domain assignments, but not the alignment regions (Pfam-Consortium 2010). Unfortunately the PTMDB is based on those annotations found in the provided flat files, which contain the envelope coordinates. There are therefore overlapping domain assignments in the PTMDB. The second mechanism is explained in the documentation that accompanies the Pfam web site. It states that nested domain annotations are allowed when it is believed that the insertion of one domain into another does not affect the tertiary structure of the domain (Pfam-Consortium 2010) (see Figure 36 for an example).

To simplify the analysis routines written for PTM Browser, all overlapping Pfam domain assignments in the PTMDB have been resolved to a single assignment.

Overlapping assignments have been resolved by keeping only the assignment whose start coordinate is closest to the modification annotation.

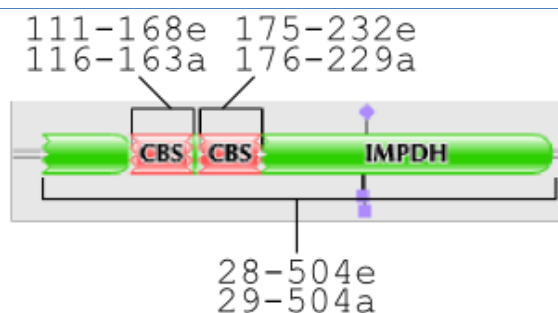


Figure 36: Domain layout of the protein IMDH1_HUMAN (SwissProt ID). Alignment coordinates are suffixed with “a” and envelope with “e”. This protein has two CBS domains which are nested in a single IMPDH domain.

Section 5.3 Conservation analysis workflows

PTM Browser is a web application that can perform either a protein family or taxonomic PTM conservation analysis. This tool has been designed to use the new PTMDB that was described in the three previous chapters. This database contains the total set of modification annotations from Swiss-Prot (v55.3) and Phospho.ELM (v7.0). The following sections outline how these two types of analysis can be performed using existing tools and how new workflows have been created by the PTM Browser tool.

Section 5.3.1 Protein family analysis

Section 5.3.1(a) Current workflow

The first step in analysing the conservation of PTMs in a family of proteins is to identify family members. Note that the term “protein family” is being used here to describe any group of related proteins (e.g. orthologues, in-paralogues, and out-paralogues). There are many different methods that can be used to obtain this list of family members. One of the most simplistic approaches is to search a protein database (e.g. UniProtKB) with a gene name. There are obvious issues with this approach caused by synonyms, missing gene names and coincidental gene names (genes with the same name that are not homologous). An alternative would be to identify family members from the results of BLAST/PSI-BLAST queries. There are also databases that specialise in identifying proteins that belong to specific families (e.g. Pfam, COG, KOG, KEGG and InParanoid).

Once a list of family members has been obtained the next step is to obtain their corresponding PTM annotations. The new interface to the UniProtKB (which was released after the PTM Browser interface was designed) makes this process far simpler than some of the more specialised databases. The UniProtKB interface can be queried for UniProtKB accessions with protein family identifiers (e.g. a Pfam accession). The UniProtKB interface allows users to select the entries they are interested in from the returned results (via a checkbox beside each entry).

Users then have two choices they can either download all the information the UniProtKB contains on the selected entries (which includes PTM annotations) or they can perform a multiple sequence alignment. Getting to this step is slightly different if the user has a list of UniProtKB accessions (rather than a protein family accession). To download all the information that the UniProtKB contains for the corresponding entries they must enter the accessions into the "retrieve query box". To perform a multiple sequence alignment the accessions must be entered into the "alignment query box".

By performing a multiple sequence alignment in the UniProtKB interface users can choose to highlight residues with known modification sites. It is then up to the user to deduce which modifications are conserved. If a user wishes to know what PTM type(s) a particular highlighted residue is annotated with, they must manually calculate the residue number (i.e. by counting in the alignment) and then refer back to the original UniProtKB web page for the corresponding entry.

The alternate route is for users to manually extract the PTM annotations from the downloaded UniProtKB flat file (specific to the accessions they requested annotations for) - for large protein families this will almost certainly require a small amount of programming. The benefit of downloading the annotations is that users can merge the UniProtKB annotations with those downloaded from other sources. Note that most other PTM databases will require each UniProtKB accession to be entered separately. Of course the main disadvantage is that the user is now responsible for both creating an alignment and highlighting which residues are modified.

It should be pointed out that the user must also be careful to check whether they believe that a missing annotation truly represents the likely absence of a modification. For example if the protein family contains entries from the TrEMBL database they will by definition have no PTM annotations in the UniProtKB, although they could of course have annotations in other databases (e.g. Phospho.ELM).

PhosphoSitePlus (<<http://www.phosphosite.org>>) is interesting because it explicitly provides users with information regarding the conservation of modifications (predominately for phosphorylation and acetylation PTM classes) for a specific protein. Conservation results are shown in a simple table structure. Note that the majority of PTM annotations in this database are associated with *H. sapiens*, *M. musculus* and *R. norvegicus*. This does limit the scope of any conservation analysis performed with this database to these three species, however this database does appear to contain many PTM annotations not found in Swiss-Prot. The conservation results generated directly by this database can only be used to analyse the conservation between orthologues. For example the database does not allow for a direct comparison of the p53 family of proteins, although it does contain separate entries for p53, p63 and p73 (see Section 6.2 for more information on PTM conservation in the p53 family of proteins).

Section 5.3.1(b) *PTM Browser workflow*

The PTM Browser workflow that has been created to analyse protein families is shown in Figure 37. Users can either supply a predefined list of proteins, or provide a single protein (see Figure 38) and request that a new protein list is created from the orthologues present in the CoPaO dataset. Note that the second route is limited to proteins from species that were included in the CoPaO analysis presented in Chapter 3 (30 species). Users can restrict the CoPaO route analysis to only include proteins from a pair of specific species - otherwise all orthologues are added to the protein list.

The user defined protein list is restricted to proteins that were present in the cross-annotation query set (see Chapter 4) and have had Pfam annotations

imported into the PTMDB. This restriction has been put in place to prevent the display of misleading results.

Once the protein list has been defined the corresponding modification annotations are extracted from the PTMDB. Note that at present only those annotations that fall in a Pfam domain are accepted. Figure 37 shows that the next step is to overlay Pfam domain alignment columns onto the initial modification dataset (indexed by amino acid residue number). In reality this information is permanently stored in the PTMDB alongside each annotation (see Section 5.2.1).

The Pfam indexed modification dataset is then converted into a number of different formats. A conservation table is generated that lists the modifications present in the protein family (by Pfam domain accession, alignment position and PTM class) against the members of the protein family (listing if the modification is present and at which residue(s)). This table is produced in both an Excel and HTML format (see Figure 39 for an example). This conservation table addresses the previously raised issue with the proposed UniProtKB workflow, whereby a user had to manually calculate residue numbers and manually identify the modifications present.

For each Pfam domain that is present in the protein family a separate starred alignment is created (see Figure 40). Note that a multiple sequence alignment of each domain is not performed as those alignments produced by the Pfam project are stored in the PTMDB. These stored alignments are used to create a new alignment which only contains those proteins in the original protein list. Residues that have modification annotations are starred in these alignments.

Finally the Pfam indexed modification dataset is exported to an XML and SQL dump file – predominantly aimed at users of the web service API.

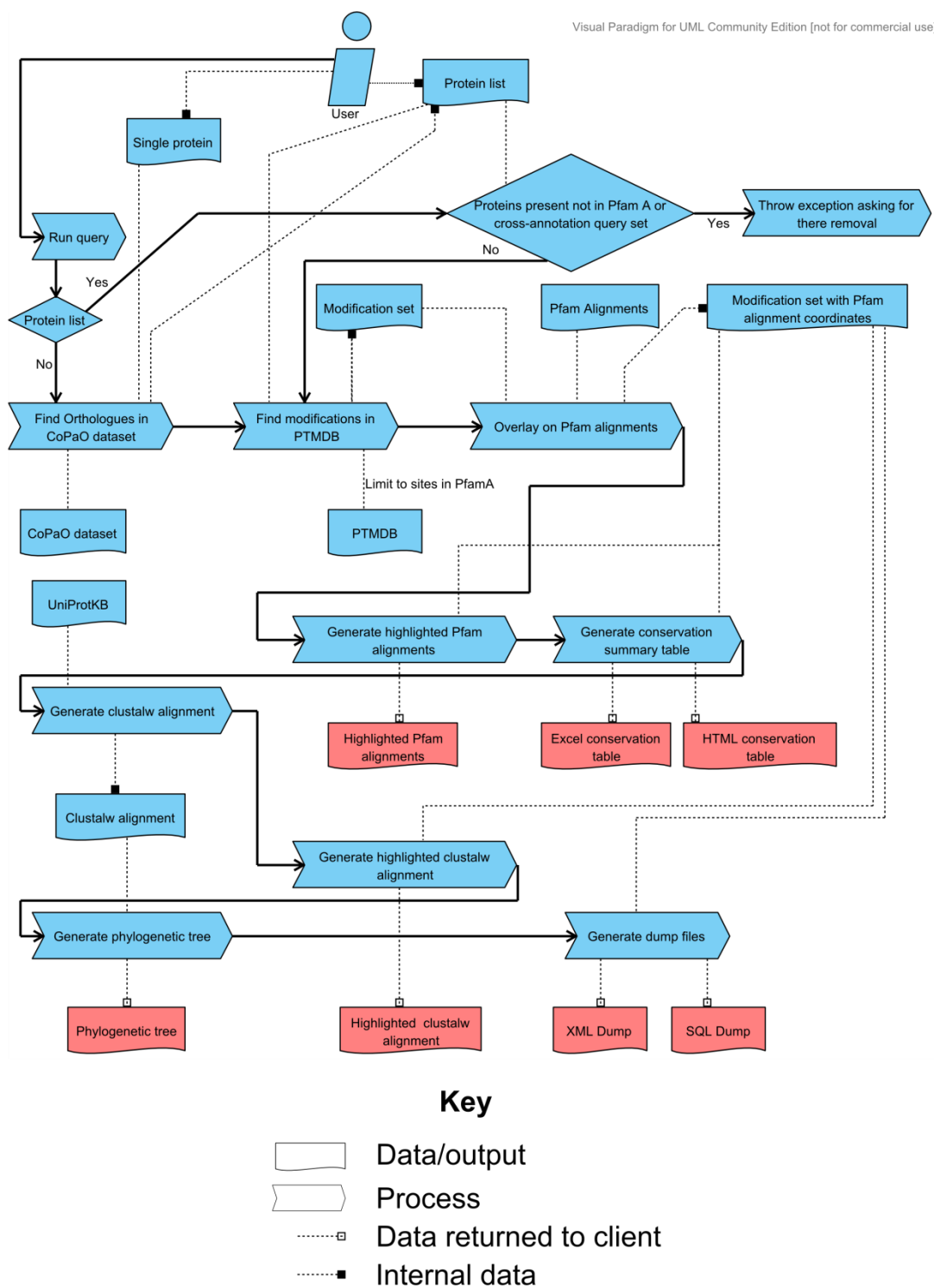


Figure 37: PTM Browser protein family workflow.

Figure 38: PTM Browser accession entry component.

This component can be used to enter either a single UniProtKB entry or a list (space or comma separated) of entries. A conservation analysis can be performed on this list with the “Show Conservation” button.

PTM Class	PfamAC	PfamPos	O09185	P07193	P67938
Phosphoprotein	PF00870	4	99 ✓	73 ✓	91 ✓
Phosphoprotein	PF00870	171	215 ✓	190 ✓	208 ✓
Phosphoprotein	PF08563	1	6 ✓	5 ✓	✗
Phosphoprotein	PF08563	4	9 ✓	✗	✗
Phosphoprotein	PF08563	10	15 ✓	14 ✓	15 ✓
Phosphoprotein	PF08563	13	18 ✓	17 ✓	18 ✓
Phosphoprotein	PF08563	15	20 ✓	✗	20 ✓

Figure 39: PTM Browser conservation table format

To aid the user in understanding the evolutionary relationships that exist between the proteins in the original protein list, a phylogenetic tree is produced from a full length multiple sequence alignment. The multiple sequence alignment is produced by CLUSTAL W (Thompson, Higgins, and Gibson 1994) and the phylogenetic tree by Phylip (Felsenstein 1989). Modification positions are starred in the CLUSTAL W multiple sequence alignment. It is extremely important to note that the CLUSTAL W alignment will not necessarily match the Pfam domain alignments in the corresponding regions. This can lead to an increase in the number of alignment columns being starred in one alignment versus the other. Therefore the CLUSTAL W alignment may indicate that different residues are homologous to each other than those indicated by the Pfam domain alignments. Note that the conservation summary table always

agrees with the Pfam domain alignments and not that generated by CLUSTAL W alignment.

```
|O09185|Cricetulus griseus [Chinese hamster]      Q*SDL*STELPL*SQE*TF*SDLWKLLPPN
|P07193|Xenopus laevis [African clawed frog]      S*SET~GMDPPL*SQE*TF~EDLWSLLPDP
|P67938|Bos indicus []                            Q~AEL~NVEPPL*SQE*TF*SDLWNLLPEN
|P67939|Bos taurus [cattle]                      Q~AEL~NVEPPL*SQE*TF*SDLWNLLPEN
|Q29537|Canis lupus familiaris [dog]              Q*SEL~NIDPPL*SQE*TF*SELWNLLPEN
```

Figure 40: PTM Browser Pfam domain starred alignment example.

Note that this alignment is for the p53 transcriptional activation domain - PF08563.

An example of the format that results are returned to users conducting a protein family analysis is shown in Figure 41.

PTM Class	P67938	Q9W678
Phosphoprotein	✓	✓

PTM Class	PfamAC	PfamPos	P67938	Q9W678
Phosphoprotein	PF00870	4	✓	✗
Phosphoprotein	PF00870	171	✓	✓
Phosphoprotein	PF08563	10	✓	✗
Phosphoprotein	PF08563	13	✓	✗
Phosphoprotein	PF08563	15	✓	✗

Figure 41: PTM Browser conservation results panel.

Section 5.3.2 Taxonomic comparisons

Section 5.3.2(a) Current workflow

With current resources it is nearly impossible for end-users to carry out a global conservation analysis between species let alone taxonomic groups (e.g. Eukaryotes and Bacteria). The only exception is for species that have very few PTM annotations – which makes manual inspection and analysis possible. Large scale analysis between species is likely to prove most valuable to end-users in identifying potential new avenues of research, without the requirement to select a protein family of interest.

Section 5.3.2(b) PTM Browser workflow

A simplified version of the PTM Browser taxonomic comparison workflow is shown in Figure 42. The taxonomic workflow has been implemented in PTM Browser on top of a more generic workflow that enables users to compare PTM

annotations from two collections (or groups of annotations) that they define. Using the taxonomic workflow users select which species they would like to be present in each of the two groups (the simplest analysis would involve one species in each group). In PTM Browser this workflow is able to accept any node from the NCBI Entrez Taxonomy, thus enabling the comparison of mammals and amphibians for instance. By default PTM Browser will analyse the conservation of modifications between the two groups based on Pfam domain alignments. Alternatively users can demand that an annotation is only conserved when it is found in the corresponding orthologue from the CoPaO dataset. Forcing the conservation analysis to be carried out between orthologues, rather than inside a Pfam domain, restricts the analysis to the 30 species present in the CoPaO dataset (and makes it impossible to carry out an analysis for taxonomic groups, such as mammals). Note that when PTM Browser compares annotations between orthologues, Pfam domain alignments are still used to identify homologous residues.

XML and SQL dumps are produced that contain the conserved PTM annotations. In addition a scrollable HTML results page as well as summary graphs and tables are produced.

The next section describes in more detail the generic workflow that has been created in PTM Browser that can be used to compare modifications between species.

Section 5.3.3 Generic conservation workflow

The generic workflow is centred on the idea of placing PTM datasets, called experiments, into two groups. Users cannot currently define their own PTM annotation datasets directly (i.e. with annotations from their own database of PTMs) – although this is being considered as a future addition to the current workflow. Instead users define parameters that are used to populate a PTM dataset with annotations from the PTMDB.

Section 5.3.3(a) *Defining PTM datasets*

An experiment has the following six query attributes that determine which annotations are included in a set: PTMDB subset (PTMDB_{Full}, PTMDB_{PfamA}),

origin database (Swiss-Prot, Phospho.ELM, spPfamA, and tremblPfamA), taxonomic node, evidence qualifier, PTM class, and PTM type. Note that spPfamA and tremblPfamA contain all those cross-annotations made to proteins in Swiss-Prot and TrEMBL databases, respectively.

Figure 43 shows the PTM Browser “Add Experiment” component, which is used to select the desired values for each of the six search parameters an individual experiment contains. All attributes are selected using the provided dropdown menus except for the taxonomic node. Note that users must choose a PTM class before they can select a PTM type.

To aid the user in their selection of a valid value for the species field, possible selectable matches in the PTMDB are displayed to the user as they type. The species attribute of an experiment must contain either a valid ID or scientific name from the NCBI Entrez Taxonomy (NET). The NCBI Entrez taxonomy web site has been embedded in PTM Browser, to allow users to search for taxa they would like to investigate. PTM Browser supports the notation of a directly and indirectly annotated taxon. For example *H. sapiens* is a directly annotated taxon because there are UniProtKB entries tagged with this species name. In contrast Mammalia is an indirectly annotated taxon as UniProtKB does not contain any entries tagged with this taxon name, but rather taxa that descend from this node in the NCBI Entrez Taxonomy tree (e.g. *M. musculus*). To include all annotations indirectly annotated to descendent taxa, the provided checkbox “Include child nodes” must be selected. Note that PTM Browser does not check whether a taxon selected by a user is directly, or indirectly, annotated - therefore it cannot enable/disable this option automatically.

Once all attributes have been selected the user simply clicks the plus icon, which inserts the experiment into the experiment browser - this is discussed shortly.

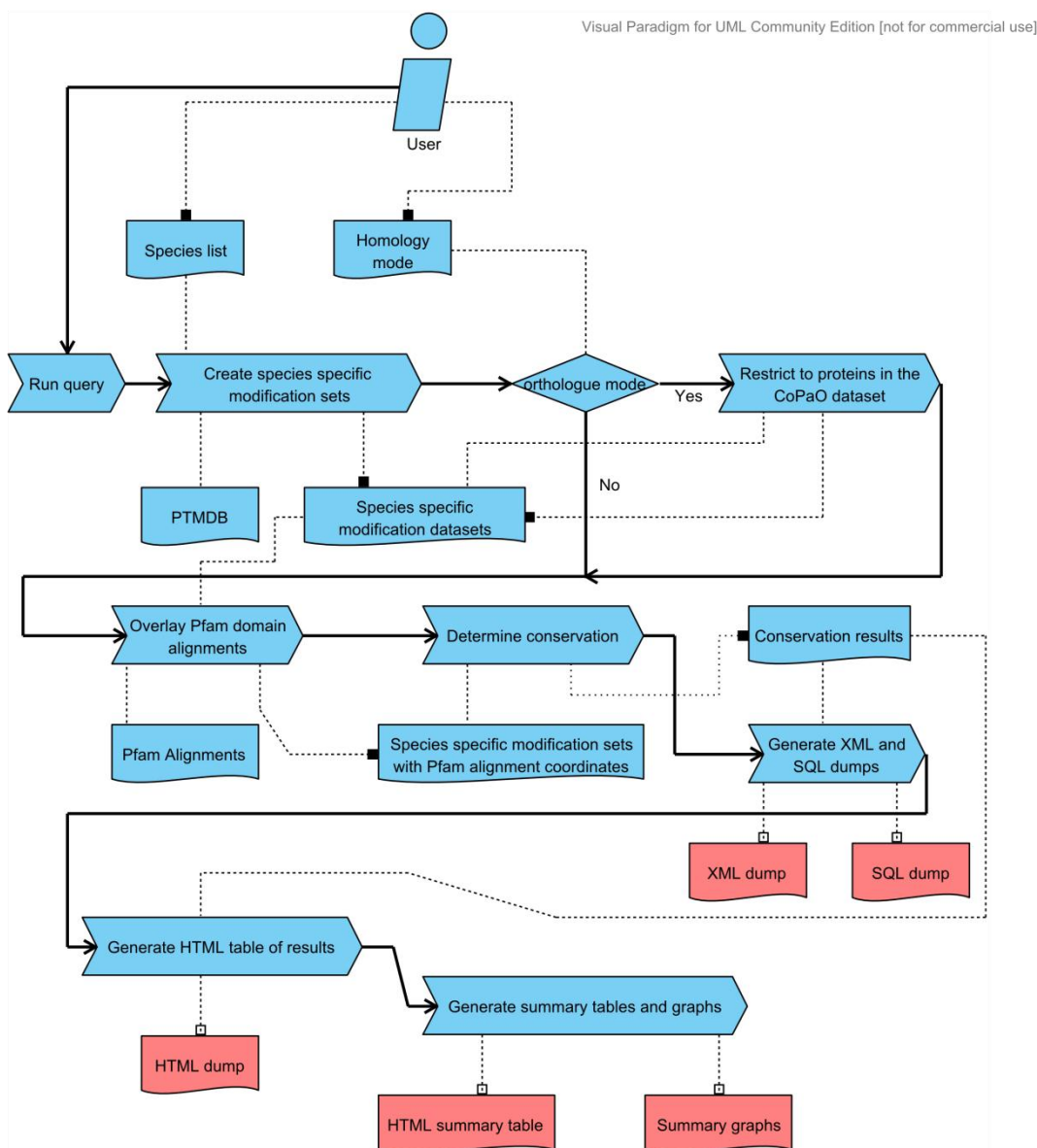


Figure 42: PTM Browser taxonomic comparison workflow.

The server side component of PTM Browser creates all PTM annotation datasets as new MySQL tables. To prevent the same dataset being created multiple times (either for the same user or different users) this component keeps a record of the query parameters used to create a new table. One benefit of storing the datasets as MySQL tables is that PTM Browser can take advantage of the built-in indexing facilities of MySQL to speed up queries involving multiple datasets. Note that the server side component has been programmed to automatically index these tables as required.

(a)

(b)

(c)

(d)

Figure 43: (a) User interface to add experiments to the experiment browser panel. (b) Auto-completion of a directly annotated taxon. (c) Auto-completion of an NCBI Entrez Taxonomy identifier. (d) Auto-completion of an indirectly annotated taxon.

Section 5.3.3(b) Grouping PTM datasets

As previously stated this workflow requires that users place PTM datasets into two groups. Requiring PTM datasets to be placed into two groups may seem unnecessary when only comparing the annotations of two species. However it becomes useful when a user wishes to run more complex queries. For example

a user may want to know which modification annotations are present in *E. coli* and either *H. sapiens* or *M. musculus*. This is easy to accomplish with PTM Browser, by placing the *E. coli* dataset into one group and those of *H. sapiens* and *M. musculus* into the other. More examples are provided throughout this chapter of placing multiple PTM datasets into the same group.

When a PTM dataset is created by the user they are automatically added to the experiment browser component. The experiment browser is located in the left panel of PTM Browser. Figure 50 (a) shows an image of the experiment browser after four experiments had been created. This component allows users to review the experiments they have created as well as carry out the following actions: grouping, deletion and application of additional constraints (which are described in Section 5.3.3(e)). Left-clicking on an experiment selects it and activates many of the actions just described. Figure 50 (b) and (c) demonstrate the process of selecting and grouping experiments. Note that the two groups of experiments are highlighted in two separate colours.

(a)

Species	Child nodes	Database	Evidence	Class	Type	Exclude	Restrict	Set	ID
Homo sapiens	no	Any	Any	Glycosylation	Any	<input type="radio"/>	<input type="radio"/>	PfamA	0
Homo sapiens	no	Any	Any	Phosphoprotein	Any	<input type="radio"/>	<input type="radio"/>	PfamA	1
Mus musculus	no	Any	Any	Glycosylation	Any	<input type="radio"/>	<input type="radio"/>	PfamA	2
Mus musculus	no	Any	Any	Phosphoprotein	Any	<input type="radio"/>	<input type="radio"/>	PfamA	3

(b)

Species	Child nodes	Database	Evidence	Class	Type	Exclude	Restrict	Set	ID
Homo sapiens	no	Any	Any	Glycosylation	Any	<input type="radio"/>	<input type="radio"/>	PfamA	0
Homo sapiens	no	Any	Any	Phosphoprotein	Any	<input type="radio"/>	<input type="radio"/>	PfamA	1
Mus musculus	no	Any	Any	Glycosylation	Any	<input type="radio"/>	<input type="radio"/>	PfamA	2
Mus musculus	no	Any	Any	Phosphoprotein	Any	<input type="radio"/>	<input type="radio"/>	PfamA	3

Figure 44: PTM Browser experiment component. (a) showing two selected experiments, (b) view after group action has been carried out.

Section 5.3.3(c) *Conservation analysis types*

This workflow supports both an intersection and complementation conservation analysis. The first, intersection, can be used to return only those annotations that are found in both groups of PTM annotations. The second, complementation, does the opposite by returning only those PTM annotations that are present in group one and not group two. A limitation of the complementation analysis is that it must be run twice if the user wishes to know which annotations are unique to each group. This requires the grouping to be manually inverted by the user (i.e. group one, becomes group two).

Figure 45: Query selection panel. Query mode can be selected as union, intersect and complement. When intersect or complement are selected the field constraints can be selected. When the user has entered orthologue/parologue mode the user is able to select whether they wish to retrieve annotations for orthologues, paralogues or both.

The fields that are used to identify matching annotations are controlled by the user. Users can select any combination of the following fields Pfam domain, Pfam alignment position, PTM class, PTM type, UniProtKB accession and primary sequence residue number. The final two fields (UniProtKB accession and residue number) are not relevant when comparing PTM datasets from different species. In PTM Browser users select the conservation mode (intersect or complement) and fields to match against, using the “Query selection panel” shown in Figure 45.

The server side component of PTM Browser begins a taxonomic conservation analysis by first creating each PTM annotation dataset. The annotations from each dataset are then merged into corresponding group PTM datasets (represented by new MySQL tables). Following this, two new “unique value” MySQL tables are created, one for each group, which include all unique values observed for the user selected fields in the corresponding merged group table. For example consider the small PTM annotation dataset shown in Table 39. If a user had requested an intersect between this dataset and another using the fields Pfam accession and Pfam alignment position – the contents of the corresponding “unique value” table would match that shown in Table 40. Finally a third new MySQL table is created, which either contains all rows present in both “unique value” tables (for an intersection query) or those present in the group one “unique value” table not present in that for group two (for a complementation query). Note that throughout this process a record is kept of which new tables have been created and the arguments that have been used (for example the “unique value” tables) – therefore were possible PTM Browser will use existing tables to process a user conservation query.

Accession	Start	PTM Class	PTM Type	Pfam Accession	Aln Position
P05813	82	Methylation	S-methylcysteine	PF00030	120
P07315	23	Methylation	S-methylcysteine	PF00030	39
Q5VXM1	348	Methylation	Methylhistidine	PF00431	689
Q9BY79	389	Methylation	Methylhistidine	PF00431	689
O75367	18	Methylation	N6-methyllysine	PF00125	3
P62807	47	Methylation	N6-methyllysine	PF00125	19
P58876	47	Methylation	N6-methyllysine	PF00125	19

Table 39: Example annotation set for an experiment.

Pfam Accession	Aln Position
PF00030	120
PF00030	39
PF00431	689
PF00125	3
PF00125	19

Table 40: List of the unique combinations of the fields Pfam Accession and Aln Position observed in the annotation set shown in Table 39.

Section 5.3.3(d) *Results*

The workflow returns the list of “unique values” that either matched (intersect) or did not (complement); this is equivalent to the third conservation table described in the previous section. The workflow has been designed so that users can request a list of all the raw annotations (from the merged group tables) that correspond to the entries in the “unique value” table. The PTM Browser client side component displays these annotations in the format shown in Figure 46. Note that users can also request only those annotations that match specific rows in the “unique value” table. In addition all annotation result sets are automatically converted into SQL and XML dumps.

The server side component of PTM Browser automatically generates summary statistics and graphs. The summary tables describe the degree of conservation between the two groups by the number of sites, proteins and domains – grouped by PTM class. Note that the summary graphs are produced as SVG (Support Vector Graphic) files from the summary tables. The summary tables are returned in HTML and Excel format. An example summary table for an intersection experiment is shown in Figure 47 and an example graph is shown in Figure 48.

It was considered likely that users might want to search for particular proteins in the returned intersect or complement result sets. The “accession entry” box (shown in Figure 38) can be used for this purpose – by entering the accession list and selecting “Constraint current result set”.

(a)

Summary of results			
View annotations			
Results			
Next	Save to file	Download SQL dump	Expand all 1/100 (589)
PfamAccession	AlnPosition	Expand	
PF00001	867	<input checked="" type="checkbox"/>	
PF00001	903	<input checked="" type="checkbox"/>	
PF00001	1366	<input checked="" type="checkbox"/>	
PF00001	1377	<input type="checkbox"/>	

(b)

Summary of results

View annotations

Save to file

Back

1/3 (3)

UniProtKBAccession	Start	KwDescription	PtmDescription	MethodDescription	DatabaseName	PfamAccession	AlnPosition	NetId
Q6JFG1	140	Phosphoprotein	Phosphoserine	Experimental	SwissProt	PF00001	867	9606
Q6JFG1	144	Phosphoprotein	Phosphoserine	Experimental	SwissProt	PF00001	903	9606
Q9GZQ6	216	Phosphoprotein	Phosphotyrosine	Experimental	SwissProt	PF00001	1366	9606

Show API Call

Figure 46: (a) A complement has been performed between *Homo sapiens* Phosphorylation annotations and those of *Mus musculus*. The intersection fields Pfam accession and Pfam alignment column have been selected. The *Homo sapiens* annotations were placed into group one. Therefore the shown Pfam domain positions have at least one phosphorylation annotation in the *Homo sapiens* phosphorylation annotation set and none in the *Mus musculus* set. (b) The expand button has been pressed on the panel shown in (b) which displays those annotations that are connected to those complement/intersection rows that have their checkbox ticked.

Summary of results											
count=distinct protein positions											
download as excel spreadsheet											
KwDescription	By similarity		Experimental		Potential		Probable				
	Total intersect	%	Total intersect	%	Total intersect	%	Total intersect	%	Total intersect	%	
Acetylation-1*	91	88	96.70	218	175	80.28	0	0	0	2	2 100.00
Acetylation-2*	194	191	98.45	99	91	91.92	0	0	0	5	4 80.00
ADP-ribosylation-1*	2	2	100.00	0	0	0	8	8 100.00	0	0	0
ADP-ribosylation-2*	25	19	76.00	0	0	0	8	8 100.00	2	2 100.00	
Amidation-1*	32	31	96.87	4	4	100.00	0	0	0	2	1 50.00
Amidation-2*	12	12	100.00	26	25	96.15	0	0	0	1	1 100.00
Citrullination-1*	3	3	100.00	1		0.00	0	0	0	0	0
Citrullination-2*	0	0	0	5	3	60.00	0	0	0	0	0
FAD-1*	6	5	83.33	0	0	0	0	0	0	1	1 100.00
FAD-2*	4	4	100.00	2	2	100.00	0	0	0	0	0
Flavoprotein-1*	6	5	83.33	0	0	0	0	0	0	1	1 100.00
Flavoprotein-2*	4	4	100.00	2	2	100.00	0	0	0	0	0
Gamma-carboxyglutamic acid-1*	64	64	100.00	20	20	100.00	0	0	0	0	0
Gamma-carboxyglutamic acid-2*	4	3	75.00	80	77	96.25	0	0	0	0	0

Figure 47 PTM Browser example of a summary of an intersection experiment between two species.

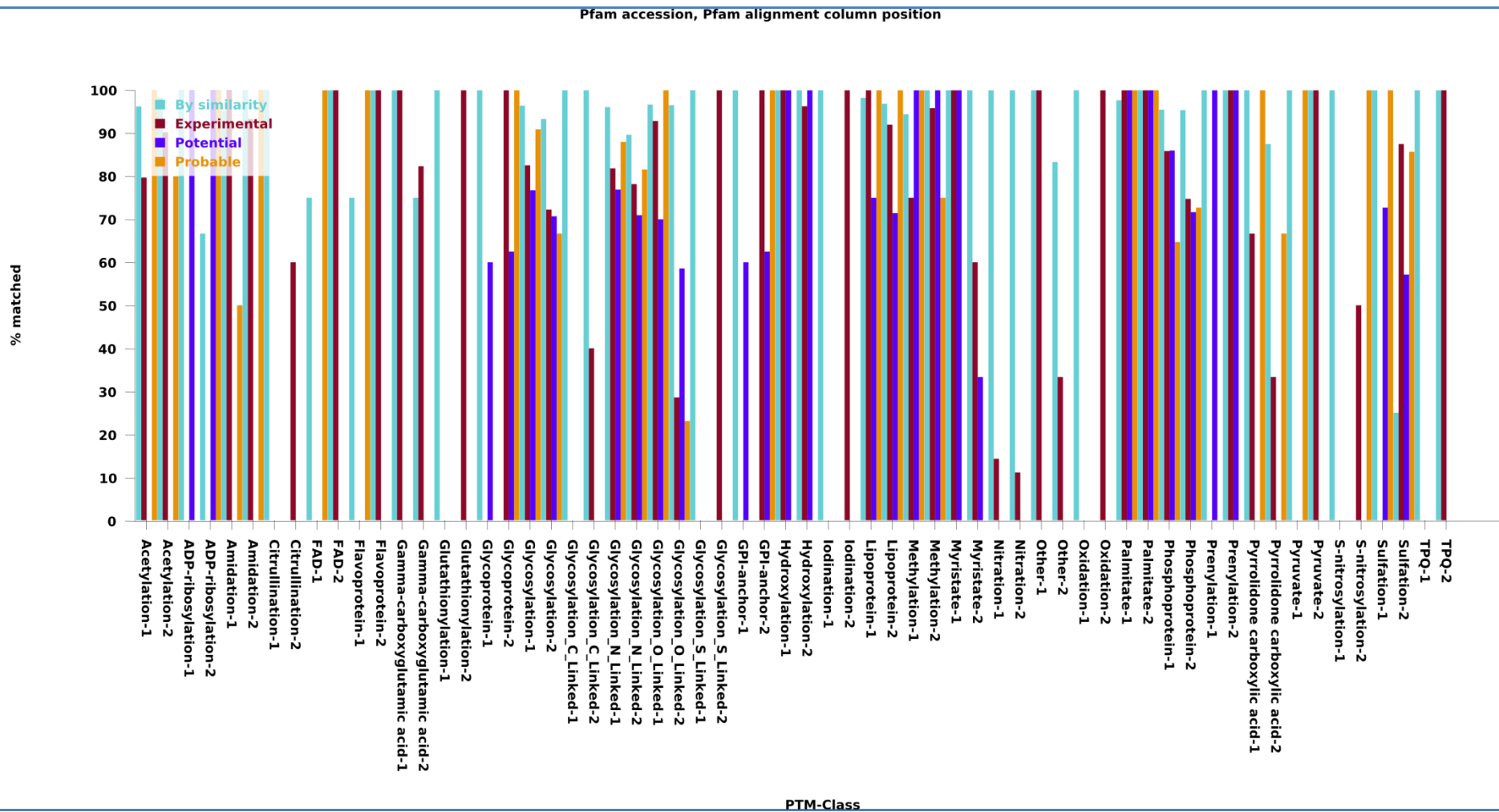


Figure 48: PTM Browser example intersection graph – shows the percentage of each modification in each group that intersects with annotations in the other group. The group number is appended to each PTM class – to show data for each group

Section 5.3.3(e) Constraints

PTM Browser allows for one experiment to be used as a constraint upon the annotations retrieved for another. The constraint can be setup to exclude or restrict the annotations in the “real experiment” that match those in the “constraint experiment”. Multiple constraints are allowed for each experiment. No attempt is made by PTM Browser to prevent overlapping conflicting constraints from being applied to the same experiment (i.e. an attempt to both exclude and restrict the same annotation). Conflicts are instead resolved on the server by allowing exclude constraints to override restrictions.

The constraint includes a list of user-chosen fields that should be used to identify matching annotations. Users can select any combination of the following fields: UniProtKB accession, residue number, Pfam accession, Pfam alignment column, PTM class and PTM type. Matching annotations are identified using the same technique discussed in Section 5.3.3(c)

There are two ways to apply constraints to experiments: i) as an intragroup constraint using existing experiments, and ii) by creating a new constraint that is applied to the selected experiments.

The first method allows for one experiment, which is part of a group, to be applied as a constraint against all the other members. Intragroup constraints can be applied by checking the required restrict or exclude checkbox found next to each experiment in the experiment browser (as shown in Figure 44).

Alternatively a new experiment can be directly applied as a constraint from the “Add experiment” panel to any experiments that are selected in the experiment browser. The “Add as constraint” checkbox must be ticked before the experiment is added for it to be applied as a constraint. Multiple constraints can be applied to each experiment and are viewed by simply selecting a single experiment in the experiment browser (see Figure 49). These constraints default to identifying matching annotations simply by the UniProtKB accession field. This can be changed with the procedure discussed below.

Experiment browser

Delete, or add constraint to selected experiment

Species	Child nodes	Database	Evidence	Class	Type	Exclude	Restrict	Set	ID
Homo sapiens	no	Any	Any	Glycosylation	Any	<input type="radio"/>	<input type="radio"/>	PfamA	0
Homo sapiens	no	Any	Any	Phosphoprotein	Any	<input type="radio"/>	<input type="radio"/>	PfamA	1
Mus musculus	no	Any	Any	Glycosylation	Any	<input type="radio"/>	<input type="radio"/>	PfamA	2
Mus musculus	no	Any	Any	Phosphoprotein	Any	<input type="radio"/>	<input type="radio"/>	PfamA	3

Run

Experiment specific constraints

Species	Child nodes	Database	Evidence	Class	Type	Exclude	Restrict	Set	ID
Homo sapiens	no	Any	Any	Methylation	Any	<input checked="" type="radio"/>	<input type="radio"/>	PfamA	0

Figure 49: Experiment browser showing one selected experiment and its associated constraint.

The fields which are used to identify matching annotations can be changed by checking the exclude or restrict checkboxes next to each experiment. Checking either box displays the dialog box shown in Figure 50. This dialog box allows the user to select which fields they would like to use to identify matching rows between experiments. Note that, as already discussed, the PTMDB_{Full} database does not contain any Pfam annotations. Therefore if either the constraint or real experiment has been set to use this database the Pfam accession and Pfam alignment column fields are not selectable.

Select join parameters

Protein space

Accession Residue position

Pfam space

Pfam Domain Pfam Position

Modification space

PTM Class PTM Type

Update

Figure 50: Dialog box presented to the user when an exclude or restrict constraint checkbox is ticked. This dialog box allows the user to select the fields over which matching annotations should be identified.

Example constraint

Exclude and restrict constraints are synonymous with the mechanism used to carry out intersect and complement queries – as discussed later. What follows is therefore only a brief account of how these constraints are applied.

A user wishes to include only those methylation sites from *M. musculus* that are also methylated in *H. sapiens* in a **single experiment**. They have also decided that they are going to identify matching sites based on Pfam accession and

Pfam alignment column position. First they create the *M. musculus*/methylation experiment. Next they select this experiment in the experiment browser and check the “Add as constraint” checkbox found on the “Add experiment” panel. They now enter all the values into the “Add experiment” panel for the *H. sapiens*/methylation experiment and click add. Finally they check the “restrict” checkbox beside the new experiment and select the two fields “Pfam accession” and “Pfam alignment column”.

Section 5.3.3(f) Restricting to orthologues

This workflow also allows users to request that annotations are matched only between orthologues. PTM Browser requires that users decide at the start of their query if they would like to run an orthologue based query. This is to allow the client software to make sure that only valid taxonomic comparisons are requested (i.e. between species in the CoPaO dataset).

The user starts by entering an experiment for the first species they wish to compare with the “Add experiment” panel. Only those species that have been compared with another using the CoPaO program are suggested to the user. After adding this experiment to the experiment browser, it is automatically added to group one. The user is then immediately prompted to enter the experiment for their second species (with which they want to compare the first). Only those species which have been compared using the CoPaO program with the species selected for the first experiment will be suggested to the user. Once this experiment is added, it is automatically added to group two.

Now that two experiments have been created and PTM Browser knows which species the user wishes to compare further experiments can be added to each group. To add another experiment to a group an experiment in the destination group must first be selected. Once it is selected the species field, present on the “Add experiment” panel, is greyed out and set to the value of the species field of the selected experiment. Clicking “add” at this point will insert the new experiment into the destination group. Note that intragroup constraints function as previously described. New experiments can be added directly as constraints to those experiments selected in the experiment browser as described before

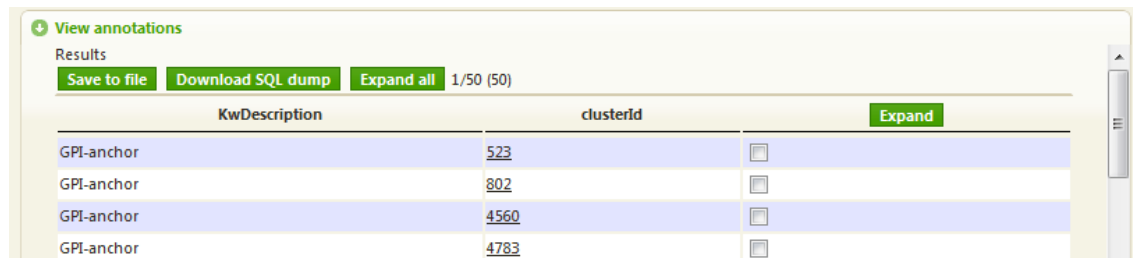
but when the “Add as constraint” checkbox is ticked you can choose any species (including indirectly annotated taxa).

As an additional option, users can select whether they wish to compare orthologues, paralogues or both (with the first and last options being most likely) using the accordion shown in Figure 45. In addition the paralogues that are included in the query can be restricted based on the InParanoid confidence values (See Section 3.3.1).

Conceptually intersects and complements in CoPaO-mode function by automatically adding the field in the PTMDB that stores the CoPaO cluster, to which a protein belongs, to the list of fields used to identify matching annotations (the constraint fields set by the user).

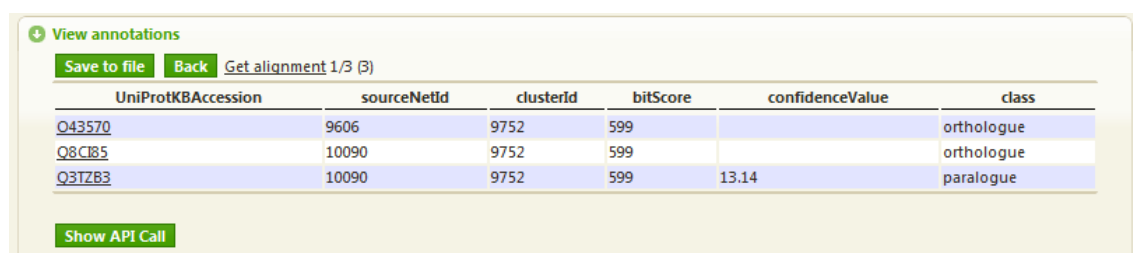
Whenever a cluster identifier is presented to a user, these can be clicked on to show further information regarding the conservation and membership of the cluster (see Figure 51).

(a)



KwDescription	clusterId	Expand
GPI-anchor	523	<input type="checkbox"/>
GPI-anchor	802	<input type="checkbox"/>
GPI-anchor	4560	<input type="checkbox"/>
GPI-anchor	4783	<input type="checkbox"/>

(b)



UniProtKBAccession	sourceNetId	clusterId	bitScore	confidenceValue	class
Q43570	9606	9752	599		orthologue
Q8CIB5	10090	9752	599		orthologue
Q3TZB3	10090	9752	599	13.14	parologue

Figure 51: Viewing a CoPaO cluster. (a) Cluster identifiers are always shown as hyperlinks (b) Clicking on a cluster identifier displays: the members of the cluster, associated class, bit score and confidence value where the class is a parologue.

Section 5.3.3(g) Constraining cross-annotations

It was previously shown that the cross-annotation dataset appears to contain a large number of annotations that are highly unlikely to be valid, based on both the number of inter-super-kingdom cross-annotations that have been observed

and the propensity of low scoring alignments. Therefore PTM Browser has been designed with the ability for the cross-annotations to be constrained based on window size and identity count (see Section 4.4.3(a)). These constraints can be entered using the dropdown menus shown in Figure 45.

PTM Browser will select all cross-annotation by default (including inter-superkingdom annotations). This decision was taken as it is very easy for users to constrain cross-annotations to only those with matching sites in the species of their choice. For example if an experiment contains all *H. sapiens* annotations (from all databases including those belonging to spPfamA and tremblPfamA cross-annotation sets) a constraint can be added that forces the removal of annotations that do not match at least one (from the Swiss-Prot or Phospho.ELM sets) belonging to a Eukaryote species (i.e. all those Eukaryote annotations that are not derived from the cross-annotated process described in Chapter 4).

Section 5.4 *Web service*

Bioinformatics developers have a great history of integrating web services into their applications and databases. Most of the major databases currently used all provide some form of programmatic access to their resources. For example the KEGG (KEGG 2010) and NCBI Entrez (NCBI-Entrez 2010) projects both provide access via the Simple Object Access Protocol (SOAP) and UniProtKB provides access via Representational State Transfer (REST). Systems have also been designed to bridge the gap between different bioinformatics web services – including the Distributed Annotation System (DAS) (see <http://www.biodas.org>) (Dowell *et al.* 2001) and BioMoby project (Wilkinson and Links 2002).

As mentioned previously the PTM Browser tool is composed of a server side and client side component. The server side component exposes a web service API for the retrieval and analysis of PTMs in the PTMDB. The client side component provides a web application for users to interact with this server side component.

A set of web services has been integrated into the PTM Browser tool that should allow bioinformaticians to access almost all of the data with which a user is usually provided when they run queries manually. This API simply involves sending POST requests to the server, which responds with the results in an XML format. Note that this web service currently requires the use of cookies, as session state has to be stored for some of the queries. To allow bioinformaticians to get started quickly with the PTM Browser API, example code is shown whenever a query is made using the PTM Browser client. The API calls can simply be shown by clicking on the “Show API” buttons that appear when queries have been made – see Figure 52. The full API is documented at the following URL <<http://wiki.ptmbrowser.org/index.php/API>>.

(a)

Hide API call

```
http://localhost:7000/processGroupRequest?group1=Homo
sapiens~undefined~Any~Any~Phosphoprotein~Any~PfamA~0~0~false~false~false~false~false~false#Homo
sapiens~undefined~Any~Any~Glycosylation~Any~PfamA~1~0~true~false~false~false~false~false,@Mus
musculus~undefined~Any~Any~Glycosylation~Any~PfamA~0~0~false~false~false~false~false~false&mode=union&
orthologueMode=false&includeOrthologues=false&includeParalogues=false
```

(b)

```
use LWP::UserAgent;
use HTTP::Cookies;
my $cookie_jar = HTTP::Cookies->new(
    file      => "cookies.lwp",
    autosave => 1,
);
my $ua = LWP::UserAgent->new;
$ua->timeout(100);
$ua->cookie_jar($cookie_jar);my $response = $ua->post('http://localhost:7000/processGroupRequest',
{group1=>"Homo
sapiens~undefined~Any~Any~Phosphoprotein~Any~PfamA~0~0~false~false~false~false~false~false#Homo
sapiens~undefined~Any~Any~Glycosylation~Any~PfamA~1~0~true~false~false~false~false~false,@Mus
musculus~undefined~Any~Any~Glycosylation~Any~PfamA~0~0~false~false~false~false~false~false",mode=>"uni
on",orthologueMode=>"false",includeOrthologues=>"false",includeParalogues=>"false",output_format=>"xml
"});
if($response->is_success){
    print $response->decoded_content;
}else{
    die $response->status_line;
}
```

Figure 52: API calls that can be used to generate the currently displayed results can always be shown by clicking the “Show API call” button. (a) Basic URL for the currently displayed results – if the URL exceeds the limit for GET requests users will need to POST the query parameters. (b) An example Perl script that can be used to retrieve the currently displayed – cookie support is included as some PTM Browser actions require state to be stored on the server

Section 5.5 *Discussion*

A tool has been presented that allows users to perform conservation and co-occurrence analyses that were previously almost impossible for non-bioinformaticians to carry out. It should be pointed out that PHOSIDA is able to carry out a similar analysis (using a similar orthology pipeline) for phosphorylation sites (Gnad *et al.* 2007). In addition UniProtKB is able to show modification sites when multiple sequence alignments are performed between entries in the said database. This tool should prove to be extremely useful for scientists that work on species that have not had the bulk of their proteins incorporated into the Swiss-Prot database (and hence have very few modification annotations), by virtue of the cross-annotations that have been made into the TrEMBL database.

PTM Browser had initially been designed around the concept that one residue in a protein could not be annotated in the PfamA boundary of multiple domain assignments. The presence of overlapping domain boundaries was only discovered late in the development of PTM Browser where they manifested themselves as inconsistencies between the results of intersect and complement queries on identical experiments. It was for this reason that the previously mentioned procedure to resolve each PTM annotation to a single PfamA domain was implemented. However there is likely to be some biologically significant information being lost with regards to, for instance, one domain being modified by virtue of a nested domain. It will therefore be important for future editions of PTM Browser to be able to correctly handle and incorporate such overlapping domain assignments.

There is a great deal of scope for PTM Browser to be further developed – many of these potential features will be discussed in Chapter 7 – the general discussion.

Chapter 6

PTM Browser in action

Section 6.1 *Summary*

In this chapter an analysis is presented on the conservation of PTMs, which has been carried out using the PTM Browser tool. An example of using PTM Browser to analyse the conservation of modifications between protein family members is first presented. This analysis looks at the conservation of modifications between members of the p53 tumour suppressor family. Members of this family are found in many distantly related eukaryotes, from *H. sapiens* to *N. vectensis*. p53 undergoes phosphorylation, methylation, acetylation, ubiquitination and sumoylation. This analysis focused on phosphorylation, as all other PTM class annotations fell out of conserved Pfam domains. The majority of bony fish and their descendants possess three p53 paralogues, p73, p63 and p53. The analysis presented in this chapter demonstrated that a core set of phosphorylation sites are conserved across all family members. In addition p53 orthologues appear to possess additional phosphorylation sites not found in the other two paralogues. Following this analysis, data is described regarding the conservation of modifications between all three super-kingdoms. According to this analysis only six modification sites are conserved across all three super-kingdoms. An analysis into the conservation of proteins between eukaryote species based on the PTM classes that they have annotations for is presented next. The results of this analysis showed that proteins with particular modifications appear to be more highly conserved than the average conservation observed between species. For example 62.48% of *H. sapiens* proteins were found to have an orthologue in *M. musculus*, compared to 77.11% of *H. sapiens* phosphoproteins. This chapter concludes with an analysis that focuses specifically on the conservation of modifications between *H. sapiens* and *M. musculus*.

Section 6.2 *Protein family analysis – p53*

In this section the following question is asked: Are phosphorylation sites conserved across p53 protein family members and can this type of analysis reveal potential novel phosphorylation sites as yet unverified? Such an analysis might for instance suggest that phosphorylation sites that were previously thought to be essential for function may not be, as they are absent in either a close or distant relative.

The analysis presented makes extensive use of the PTM Browser tool discussed previously (Chapter 4). As part of this analysis the strengths and weaknesses of the PTM Browser tool have become apparent, which may be addressed in the future.

The apoptosis control protein, p53, is regulated by multiple post-translational modifications, including phosphorylation, acetylation and ubiquitination (Brooks and Gu 2003). It has even been shown to include methylation (Chuikov *et al.* 2004), sumoylation (Rodriguez *et al.* 1999) and neddylation (Xirodimas *et al.* 2004). The p53 protein is classified as a tumour suppressor, as its loss of function, by mutation or deletion, results in tumorigenesis (Brooks and Gu 2003). Phosphorylation and acetylation of p53 are linked to its activation, and subsequent cell cycle arrest, in order to permit DNA-damage repair pathways to correct errors in DNA replication (Brooks and Gu 2003). Whereas cell cycle progression is associated with p53 ubiquitination (Brooks and Gu 2003). Phosphorylation in response to DNA damage is mediated by serine/threonine-specific protein kinases: Chk1 and Chk2 (check-point proteins) (Shieh *et al.* 2000). Ubiquitination of p53, to promote its degradation, is mediated by Mdm2, a specific E3 ubiquitin ligase (Honda, Tanaka, and Yasuda 1997).

Members of the p53 family have been identified in a wide range of eukaryotes from *H. sapiens*, to *C. elegans* and *N. vectensis* (Pankow and Bamberger 2007). Humans possess three paralogues from this family, p63, p53 and p73 (Belyi *et al.* 2010). The sea anemone (*N. vectensis*) possesses a single p53 family gene which most closely resembles a p63/p73-like gene (Pankow and Bamberger 2007). Current evidence suggests that the p53 ancestral gene was duplicated in the early vertebrate lineage; resulting in a p53-like and p63/p73-

like gene (Belyi *et al.* 2010). The bony fish are the oldest group that appear to possess all three paralogues, p63, p53 and p73 (Belyi *et al.* 2010). This suggests that a second duplication event occurred during the evolution of bony fish (Belyi *et al.* 2010).

p53, p63 and p73 all contain a transcriptional activation domain (TA), DNA binding domain and an oligomerization domain (Arrowsmith 1999) (See Figure 53 for the arrangement). In addition p63 and p73 contain an extra c-terminal SAM domain (Sterile α motif domain) (Arrowsmith 1999).

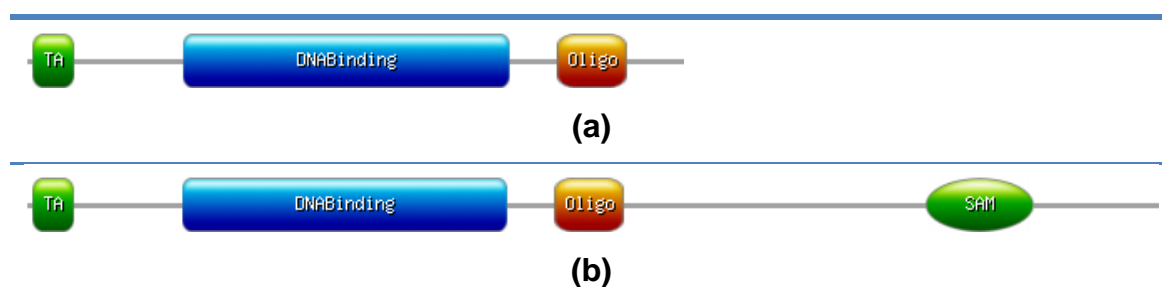


Figure 53: Domain structure of an idealised p53 (a) and p63 (b) protein. TA=Transcriptional activation domain, DNA Binding=DNA binding domain, Oligo=Oligomerization domain, SAM=Sterile alpha motif. Images produced using the mydomains tool <http://www.expasy.ch/cgi-bin/prosite/mydomains/>.

The p53 family members were identified in the PTMDB sequence space by gene name and orthologous relationships to previously identified p53 family members in the CoPaO dataset (see Chapter 3). Those proteins that were not subjected to the cross-annotation procedure (see Chapter 4) were removed from the list of p53 family members identified. In total 42 proteins were identified originating from 32 different species, of these 31 proteins were identified as p53, six as p63 and four as p73. The final protein identified in this way belonged to *N. vectensis*, which, as mentioned previously, has been characterised as a p63/p73-like protein. See Appendix 1 for the list of corresponding UniProtKB accessions. The PTM Browser tool was then used to identify which modification positions were conserved amongst the p53 family proteins. The automatically generated phylogenetic tree that PTM Browser creates is shown in Figure 54. This tree clusters the p63 and p73 genes together, and separates the p53 proteins into those from bony fish and those from mammals. Table 41 shows which PTM sites are present and which are conserved in the list of p53 family members. The PTM sites identified lie in two Pfam domains PF00870 (DNA binding domain) and PF08563 (TA domain).

The conservation of modifications across each of these domains will now be discussed separately.

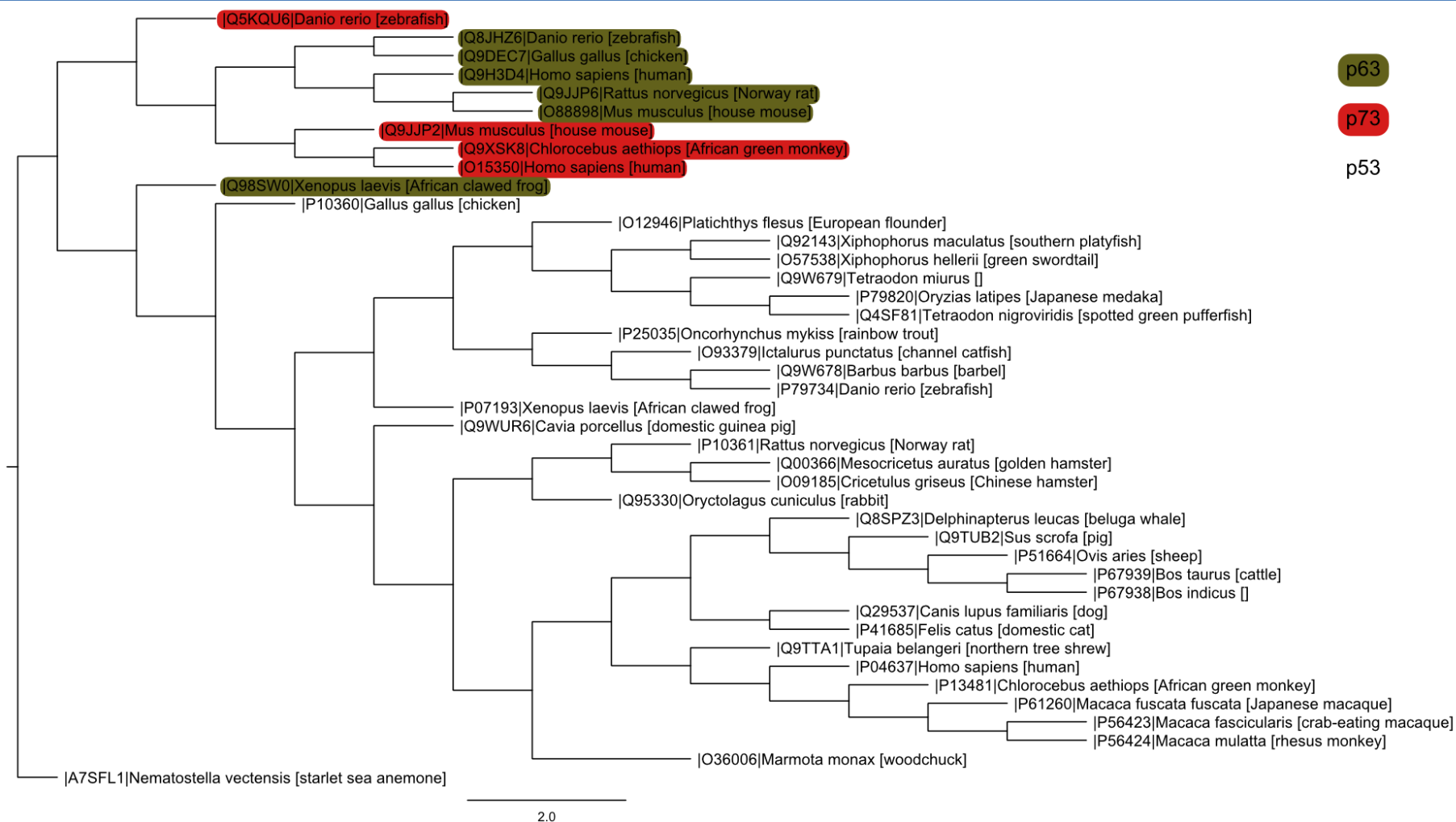


Figure 54: Phylogenetic tree showing one possible set of ancestral relationships between those proteins in the PTMDB identified as being part of the p53 family (p63, p53 and p73). This tree was produced by the PTM Browser tool using the Phylip program protpars; graphical version of this tree was produced by the FigTree program (<http://tree.bio.ed.ac.uk/software/figtree/>). Members of this family were identified by gene name and orthologous relationships in the CoPaO dataset.

Pfam accession	PF00870		PF08563					
Alignment position	4	171	1	4	10	13	15	
p63/p73 A7SFL1 Nematostella vectensis [starlet sea anemone]	✓	✓	✗	✗	✗	✗	✗	
p63 Q8JHZ6 Danio rerio [zebrafish]	✓	✓	✗	✗	✗	✗	✗	
p63 Q98SW0 Xenopus laevis [African clawed frog]	✓	✓	✗	✗	✗	✗	✗	
p63 Q9DEC7 Gallus gallus [chicken]	✓	✓	✗	✗	✗	✗	✗	
p63 Q9H3D4 Homo sapiens [human]	✓	✓	✗	✗	✗	✗	✗	
p63 Q9JJP6 Rattus norvegicus [Norway rat]	✓	✓	✗	✗	✗	✗	✗	
p63 O88898 Mus musculus [house mouse]	✓	✓	✗	✗	✗	✗	✗	
p73 Q5KQU6 Danio rerio [zebrafish]	✓	✓	✗	✗	✗	✗	✗	
p73 Q9JJP2 Mus musculus [house mouse]	✓	✓	✗	✗	✗	✗	✗	
p73 Q9XSK8 Chlorocebus aethiops [African green monkey]	✓	✓	✗	✗	✗	✗	✗	
p73 O15350 Homo sapiens [human]	✓	✓	✗	✗	✗	✗	✗	
Q92143 Xiphophorus maculatus [southern platyfish]	✗	✓	✗	✗	✓	✓	✓	
O57538 Xiphophorus hellerii [green swordtail]	✗	✓	✗	✗	✓	✓	✓	
P79820 Oryzias latipes [Japanese medaka]	✗	✓	✗	✗	✗	✗	✗	
Q4SF81 Tetraodon nigroviridis [spotted green pufferfish]	✗	✓	✗	✓	✓	✓	✗	
Q9W679 Tetraodon miurus []	✗	✓	✗	✗	✓	✓	✗	
O12946 Platichthys flesus [European flounder]	✗	✓	✗	✗	✗	✗	✗	
P25035 Oncorhynchus mykiss [rainbow trout]	✓	✓	✗	✗	✓	✓	✗	
O93379 Ictalurus punctatus [channel catfish]	✗	✓	✗	✗	✗	✗	✗	
Q9W678 Barbus barbus [barbel]	✗	✓	✗	✗	✗	✗	✗	
P79734 Danio rerio [zebrafish]	✗	✓	✗	✗	✗	✗	✗	
P07193 Xenopus laevis [African clawed frog]	✓	✓	✓	✗	✓	✓	✗	
P10360 Gallus gallus [chicken]	✓	✓	✗	✗	✗	✗	✗	
Q9WUR6 Cavia porcellus [domestic guinea pig]	✓	✓	✓	✓	✓	✓	✓	
O09185 Cricetulus griseus [Chinese hamster]	✓	✓	✓	✓	✓	✓	✓	
Q00366 Mesocricetus auratus [golden hamster]	✓	✓	✓	✓	✓	✓	✓	
P10361 Rattus norvegicus [Norway rat]	✓	✓	✓	✓	✓	✓	✓	
Q9TUB2 Sus scrofa [pig]	✓	✓	✓	✗	✓	✓	✓	
Q8SPZ3 Delphinapterus leucas [beluga whale]	✓	✓	✗	✗	✓	✓	✓	
P51664 Ovis aries [sheep]	✓	✓	✗	✗	✓	✓	✓	
P67939 Bos taurus [cattle]	✓	✓	✗	✗	✓	✓	✓	
P67938 Bos indicus []	✓	✓	✗	✗	✓	✓	✓	
Q29537 Canis lupus familiaris [dog]	✓	✓	✓	✗	✓	✓	✓	
P41685 Felis catus [domestic cat]	✓	✓	✗	✓	✓	✓	✓	
Q9TTA1 Tupaia belangeri [northern tree shrew]	✓	✓	✓	✓	✓	✓	✓	
P04637 Homo sapiens [human]	✓	✓	✓	✓	✓	✓	✓	
P13481 Chlorocebus aethiops [African green monkey]	✓	✓	✓	✓	✓	✓	✓	
P61260 Macaca fuscata fuscata [Japanese macaque]	✓	✓	✓	✓	✓	✓	✓	
P56423 Macaca fascicularis [crab-eating macaque]	✓	✓	✓	✓	✓	✓	✓	
P56424 Macaca mulatta [rhesus monkey]	✓	✓	✓	✓	✓	✓	✓	
Q95330 Oryctolagus cuniculus [rabbit]	✓	✓	✓	✓	✓	✓	✓	
O36006 Marmota monax [woodchuck]	✓	✓	✓	✓	✓	✓	✓	

Table 41: The conservation of PTMs in the p53 TA and DNA binding domains. Table was generated from the Excel file produced by the PTM Browser tool.

Section 6.2.1 p53 DNA binding domain (PF00870)

All members of the p53 family identified in the PTMDB dataset contain the p53 DNA binding domain (PF00870). The PTMDB contains annotations for two positions which are modified in this domain: 99(4) and 215(171).

Serine 99(4) is found in all p53 homologues except those belonging to the bony fish (with the exception of rainbow trout). Figure 55 shows an alignment of the region surrounding this residue. However despite being present in the Swiss-Prot database, and therefore having the potential for PTM annotations, none of the p63/p73 homologues have a phosphorylation annotation at this position. The only PTM annotation for either of these proteins in the Swiss-Prot database is for a phosphorylation on Tyr-99 of the human p73 (O15350).

The cross-annotation procedure has however created annotations at this position for this group of proteins in the PTMDB. The alignment shows that the p63/p73 sites are more similar to each other than they are to sites from p53 homologues. No literature references could be found that referred to the phosphorylation of this position for p63/p73 homologues. The PhosphoSite (Hornbeck *et al.* 2004) database (<<http://www.phosphosite.org>>) contains phosphorylation annotations for the Human, Mouse, and Rat p63/p73 homologues but not for this particular position.

The PhosphoMotif_finder (Amanchy *et al.* 2007) tool (<http://www.hprd.org/PhosphoMotif_finder>) was used to identify potential phosphorylation sites in the Human p63 /p73 homologues; based on known kinase recognition motifs. This tool matched the p63 and p73 residue 99(4) site to that recognised by casein kinase II, the p73 site was also matched to that recognised by pyruvate dehydrogenase kinase. In comparison the same site from Human p53 was found to match the motifs of the following enzymes: DNA dependent kinase, ATM kinase, PKA kinase, and PKC kinase. There is some evidence that the Human p53 is phosphorylated by the ATM kinase (Matsuoka *et al.* 2007). However, a literature search revealed no evidence that p63/p73 have been observed to be phosphorylated by casein kinase II. In contrast casein kinase II has been reported to be able to phosphorylate p53 (Filhol *et al.* 1992).

p63-p73 A7SFL1 Nematostella vectensis [starlet sea anemone]	VIAP*SSDEVPGEYSF
p63 O88898 Mus musculus [house mouse]	PAIP*SNTDYPGPHSF
p63 Q9H3D4 Homo sapiens [human]	PAIP*SNTDYPGPHSF
p63 Q8JHZ6 Danio rerio [zebrafish]	PAIP*SNTDYAGPHTF
p63 Q98SW0 Xenopus laevis [African clawed frog]	PAIP*SNTDYPGPHSF
p63 Q9DEC7 Gallus gallus [chicken]	PAIP*SNTDYPGPHSF
p63 Q9JJP6 Rattus norvegicus [Norway rat]	PAIP*SNTDYPGPHSF
p73 Q9XSK8 Chlorocebus aethiops [African green monkey]	PVIP*SNTDYPGPHHF
p73 O15350 Homo sapiens [human]	PVIP*SNTDYPGPHHF
p73 Q5KQU6 Danio rerio [zebrafish]	PAIP*SNTDYPGPHNF
p73 Q9JJP2 Mus musculus [house mouse]	PVIP*SNTDYPGPHHF
p53 Q9WUR6 Cavia porcellus [domestic guinea pig]	SSVP*SHKPYRGSYGF
p53 P04637 Homo sapiens [human]	SSVP*SQKTYQGSYGF
p53 Q9TTA1 Tupaia belangeri [northern tree shrew]	SSVP*SQKTYQGSYGF
p53 P13481 Chlorocebus aethiops [African green monkey]	SSVP*SQKTYHGSYGF
p53 P56423 Macaca fascicularis [crab-eating macaque]	SSVP*SQKTYHGSYGF
p53 P56424 Macaca mulatta [rhesus monkey]	SSVP*SQKTYHGSYGF
p53 P61260 Macaca fuscata fuscata [Japanese macaque]	SSVP*SQKTYHGSYGF
p53 P10361 Rattus norvegicus [Norway rat]	SSVP*SQKTYHGSYGF
p53 O36006 Marmota monax [woodchuck]	SSVP*SQNTYPGVYGF
p53 Q00366 Mesocricetus auratus [golden hamster]	SSVP*SYKTYQGDYGF
p53 O09185 Cricetulus griseus [Chinese hamster]	SSVP*SYKTYQGDYGF
p53 Q9TUB2 Sus scrofa [pig]	SFVP*SQKTYPGSYDF
p53 Q8SPZ3 Delphinapterus leucas [beluga whale]	SFVP*SQKTYPGSYGF
p53 Q29537 Canis lupus familiaris [dog]	SSVP*SPKTYPGTYGF
p53 P41685 Felis catus [domestic cat]	SFVP*SQKTYPGAYGF
p53 Q95330 Oryctolagus cuniculus [rabbit]	SSVP*SQKTYHGSYGF
p53 P51664 Ovis aries [sheep]	SFVP*SQKTYPGNYGF
p53 P67938 Bos indicus []	SFVP*SQKTYPGNYGF
p53 P67939 Bos taurus [cattle]	SFVP*SQKTYPGNYGF
p53 P07193 Xenopus laevis [African clawed frog]	CAVP*STDDYAGKYGL
p53 P10360 Gallus gallus [chicken]	PVVP*STEDYGGDFDF
p53 P25035 Oncorhynchus mykiss [rainbow trout]	STVP*TTSDYPGALGF
p53 O93379 Ictalurus punctatus [channel catfish]	STVP~VTSDYPGLLNF
p53 O12946 Platichthys flesus [European flounder]	STVP~VTTDYPGEYGF
p53 O57538 Xiphophorus hellerii [green swordtail]	PTVP~AISNYAGEHGF
p53 P79734 Danio rerio [zebrafish]	STVP~ETSDYPGDHGF
p53 Q92143 Xiphophorus maculatus [southern platyfish]	PTVP~AISNYAGEHGF
p53 Q9W678 Barbus barbus [barbel]	ASVP~VATDYPGEHGF
p53 Q9W679 Tetraodon miurus []	PTVP~VTTDYPGEYGF
p53 Q4SF81 Tetraodon nigroviridis [spotted green pufferfish]	PTVP~VTTDHPGEYDF
p53 P79820 Oryzias latipes [Japanese medaka]	TTVP~VTTDYPGSYEL

Figure 55: Alignment of Ser-99(4) region from various members of the p53 family present in the PTMDB. Alignment was generated automatically by the PTM Browser tool from the alignment extracted from Pfam. Note that this alignment contains a subset of those sequences present in the original Pfam alignment of this domain. Under this circumstance it's possible for an individual alignment column to be a gap column for all sequences in the alignment subset; such columns have been removed from this alignment.

Serine 215(171) is present in a highly conserved region of the p53 DNA binding domain (see Figure 56), none of the p53 family aligned are missing this residue. This residue has been shown to be phosphorylated by the Aurora-A serine/threonine kinase, with homologues having been identified in many different species (Liu *et al.* 2004). Phosphorylation by Aurora-A prevents p53 from binding to DNA and is linked to the prevention of apoptosis after DNA damage (Liu *et al.* 2004).

p63-p73 A7SFL1 Nematostella vectensis [starlet sea anemone]	ERCAQSGRL*SVKIPFHV
p63 O88898 Mus musculus [house mouse]	VEDPITGRQ*SVLVPPYEP
p63 Q9H3D4 Homo sapiens [human]	VEDPITGRQ*SVLVPPYEP
p63 Q8JHZ6 Danio rerio [zebrafish]	VEDSITGRQ*SVLVPPYEP
p63 Q98SW0 Xenopus laevis [African clawed frog]	VEDPITGRQ*SVLVPPYEP
p63 Q9DEC7 Gallus gallus [chicken]	VEDPITGRQ*SVLVPPYEP
p63 Q9JJP6 Rattus norvegicus [Norway rat]	VEDPITGRQ*SVLVPPYEP
p73 Q9XSK8 Chlorocebus aethiops [African green monkey]	VDDPVTGRQ*SVVVPYEP
p73 O15350 Homo sapiens [human]	VDDPVTGRQ*SVVVPYEP
p73 Q5KQU6 Danio rerio [zebrafish]	VDDPVTGRQ*SVLVPPYEP
p73 Q9JJP2 Mus musculus [house mouse]	VDDPVTGRQ*SVVVPYEP
p53 Q9WUR6 Cavia porcellus [domestic guinea pig]	VDDRTTFRH*SVVVPYEP
p53 P04637 Homo sapiens [human]	LDDRNTFRH*SVVVPYEP
p53 Q9TTA1 Tupaia belangeri [northern tree shrew]	SDDRNTFRH*SVVVPYEP
p53 P13481 Chlorocebus aethiops [African green monkey]	SDDRNTFRH*SVVVPYEP
p53 P56423 Macaca fascicularis [crab-eating macaque]	SDDRNTFRH*SVVVPYEP
p53 P56424 Macaca mulatta [rhesus monkey]	SDDRNTFRH*SVVVPYEP
p53 P61260 Macaca fuscata fuscata [Japanese macaque]	SDDRNTFRH*SVVVPYEP
p53 P10361 Rattus norvegicus [Norway rat]	LDDRNTFRH*SVVVPYEP
p53 O36006 Marmota monax [woodchuck]	LDDRNTFRH*SVVVPYEP
p53 Q00366 Mesocricetus auratus [golden hamster]	LDDKQTFRH*SVVVPYEP
p53 O09185 Cricetulus griseus [Chinese hamster]	LDDKQTFRH*SVVVPYEP
p53 Q9TUB2 Sus scrofa [pig]	LDDRNTFRH*SVVVPYEP
p53 Q8SPZ3 Delphinapterus leucas [beluga whale]	LDDRNTFRH*SVVVPYEP
p53 Q29537 Canis lupus familiaris [dog]	LDDRNTFRH*SVVVPYEP
p53 P41685 Felis catus [domestic cat]	LDDRNTFRH*SVVVPYEP
p53 Q95330 Oryctolagus cuniculus [rabbit]	LDDRNTFRH*SVVVPYEP
p53 P51664 Ovis aries [sheep]	FDDRNTFRH*SVVVPYEP
p53 P67938 Bos indicus []	LDDRNTFRH*SVVVPYEP
p53 P67939 Bos taurus [cattle]	LDDRNTFRH*SVVVPYEP
p53 P07193 Xenopus laevis [African clawed frog]	MEDVNSGRH*SVVVPYEP
p53 P10360 Gallus gallus [chicken]	HDDETTKRH*SVVVPYEP
p53 P25035 Oncorhynchus mykiss [rainbow trout]	MEDGNTLRH*SVLVPPYEP
p53 O93379 Ictalurus punctatus [channel catfish]	QEDGNTQAH*SVVVPYEP
p53 O12946 Platichthys flesus [European flounder]	FEDPHTKRQ*SVTVPPYEP
p53 O57538 Xiphophorus hellerii [green swordtail]	FEDPNTRRH*SVTVPPYEP
p53 P79734 Danio rerio [zebrafish]	REDNITLRH*SVFVPPYEP
p53 Q92143 Xiphophorus maculatus [southern platyfish]	FEDPNTRRH*SVTVPPYEP
p53 Q9W678 Barbus barbus [barbel]	REDDVNSRH*SVVVPYEP
p53 Q9W679 Tetraodon miurus []	FEHPHTKRQ*SVTVPPYEP
p53 Q4SF81 Tetraodon nigroviridis [spotted green pufferfish]	FEDPHTKRQ*SVTVPPYEP
p53 P79820 Oryzias latipes [Japanese medaka]	FEDPYTKRQ*SVTVPPYEP

Figure 56: Alignment of Ser-215(171) region from various members of the p53 family present in the PTMDB.

Alignment was generated automatically by the PTM Browser tool from the alignment extracted from Pfam. Note that this alignment contains a subset of those sequences present in the original Pfam alignment of this domain.

Under this circumstance it's possible for an individual alignment column to be a gap column for all sequences in the alignment subset; such columns have been removed from this alignment.

Section 6.2.2 p53 Transcriptional activation domain (PF08563)

Of the 31 p53 homologues identified in the PTMDB sequence space 25 are annotated with the Pfam domain PF08563 (p53 TA domain). Five of the p53 homologues that lack this domain annotation are fish species and the final protein belongs to chicken. None of the p63 or p73 homologues have Pfam annotations for this domain, however it has been widely reported that both p63 and p73 contain TA domains. The Pfam dataset incorporated into the PTMDB was therefore manually interrogated to check for PF08563 annotations on the six p53 homologues and all of the p63 and p73 proteins - no annotations were found. To check the integrity of the data incorporated into the PTMDB the Pfam

database was searched using the following web site <<http://pfam.sanger.ac.uk/>>, which confirmed the absence of PF08563 annotations on these proteins in the Pfam database. The CLUSTAL W alignment produced by the PTM Browser tool for the p53 family proteins was manually checked for obvious signs of homology between the p63/p73 and the p53 proteins in the region assigned to the PF08563 domain. Based on the alignment the p63/p73 TA region appears to be poorly conserved with that of the p53 proteins. However residues surrounding known p53 phosphorylation sites do appear to show evidence of conservation, although no attempt has been made to quantify and validate this.

Table 41 shows that all those p53 homologues which have an annotation for this domain have at least one PTM in the PTMDB. The alignment shown in Figure 57, produced by PTM Browser, highlights the modification sites located in this domain. Five residues have modification annotations, which correspond to the following positions in the Human p53 homologue: 6(1), 9(4), 15(10), 18(13) and 20(15).

The residues at positions 9(4), 15(10) and 18(13) have Swiss-Prot phosphorylation annotations to many p53 homologues (including Human). Prior to the cross-annotation procedure only the Rat p53 homologue (P10361) had a phosphorylation annotation for 6(1), which was imported from the Phospho.ELM database (low-throughput evidence qualifier).

The cross-annotation procedure transferred this single annotation to the following species Human, African green monkey, crab-eating macaque, Japanese macaque, rhesus monkey, northern tree shrew, golden hamster, Chinese hamster, woodchuck, dog, pig, domestic guinea pig and rabbit. The Phospho.ELM website associates this site in the Rat protein to the Casein kinase I motif. To add further validation to the transferred annotations these sites were searched for kinase phosphorylation site motifs using the PhosphoMotif_finder tool. The same Casein kinase I motif was identified at this position in all but the dog and pig sites. Note that the following five mammalian species are missing a phosphorylation compatible residue at this position, sheep, cow (both *B. taurus* and *Bos indicus*), beluga whale and cat. The

African clawed frog was the most distantly related species to Human that had both a cross-annotation and a match to the casein kinase I motif. All of the fish homologues were also lacking a phosphorylation compatible residue at this position.

P04637 Homo sapiens [human]	Q*SDP*SVEPPL*SQE*TF*SDLWKLLPEN
P13481 Chlorocebus aethiops [African green monkey]	Q*SDP*SIEPPL*SQE*TF*SDLWKLLPEN
P56423 Macaca fascicularis [crab-eating macaque]	Q*SDP*SIEPPL*SQE*TF*SDLWKLLPEN
P61260 Macaca fuscata fuscata [Japanese macaque]	Q*SDP*SIEPPL*SQE*TF*SDLWKLLPEN
P56424 Macaca mulatta [rhesus monkey]	Q*SDP*SIEPPL*SQE*TF*SDLWKLLPEN
Q9TTA1 Tupaia belangeri [northern tree shrew]	Q*SDP*SVEPPL*SQE*TF*SDLWKLLPEN
Q00366 Mesocricetus auratus [golden hamster]	Q*SDL*SIELPL*SQE*TF*SDLWKLLPEN
O09185 Cricetulus griseus [Chinese hamster]	Q*SDL*SIELPL*SQE*TF*SDLWKLLPEN
O36006 Marmota monax [woodchuck]	Q*SDL*SIEPPL*SQE*TF*SDLWNLLEN
Q9WUR6 Cavia porcellus [domestic guinea pig]	H*SDL*SIEPPL*SQE*TF*SDLWKLLPEN
Q95330 Oryctolagus cuniculus [rabbit]	Q*SDL*SIEPPL*SQE*TF*SDLWKLLPEN
P10361 Rattus norvegicus [Norway rat]	Q*SDM*SIELPL*SQE*TF*SCLWKLLPPD
Q29537 Canis lupus familiaris [dog]	Q*SEL~NIDPPL*SQE*TF*SELWNLLEN
P41685 Felis catus [domestic cat]	P~LEL*TIEPPL*SQE*TF*SELWNLLEN
Q9TUB2 Sus scrofa [pig]	Q*SEL~GVEPPL*SQE*TF*SDLWKLLPEN
P07193 Xenopus laevis [African clawed frog]	S*SET~GMDPPL*SQE*TF~EDLWSLLPDP
O57538 Xiphophorus hellerii [green swordtail]	M~EEA~DLTLPL*SQD*TF~HDLWNNVFLS
P51664 Ovis aries [sheep]	Q~AEL~GVEPPL*SQE*TF*SDLWNLLEN
P67938 Bos indicus []	Q~AEL~NVEPPL*SQE*TF*SDLWNLLEN
P67939 Bos taurus [cattle]	Q~AEL~NVEPPL*SQE*TF*SDLWNLLEN
Q8SPZ3 Delphinapterus leucas [beluga whale]	Q~AEL~GVEPPL*SQE*TF*SDLWKLLPEN
P25035 Oncorhynchus mykiss [rainbow trout]	D~LAE~NVSLPL*SQE*SF~EDLWKMNLNL
Q4SF81 Tetraodon nigroviridis [spotted green pufferfish]	M~EEE*TFSLPL*SQD*TF~QDLWENVAAP
Q92143 Xiphophorus maculatus [southern platyfish]	M~EEA~DLTLPL*SQD*TF~HDLWNNVFLS
Q9W679 Tetraodon miurus []	M~EEE~NISLPL*SQD*TF~QDLWDNVSA

Figure 57: Conservation of modifications in the p53 transcriptional activation domain (PF008563). Alignment produced by the PTM Browser tool using the raw Pfam alignment incorporated into the PTMDB.

Residue 9(4) is modified in all primate and small mammal species p53 homologues. All of these species have their p53 homologues in the Swiss-Prot section of the UniProtKB and all contain a phosphorylation annotation at this position. Residue 9(4) has been shown to be phosphorylated by Casein kinase I when serine residue 6(1) is phosphorylated in response to DNA damaging agents (including UV light and Ionising Radiation (IR) (Higashimoto *et al.* 2000). The following mammals are missing the modification site at positions 9(4) and 6(1): sheep, cow (*B. taurus* and *B. indicus*) and beluga whale. The cross-annotation procedure has transferred annotations at this position to the cat p53 homologue. However note that instead of a serine the cat homologue has a threonine at this position. The PhosphoMotif_finder tool was used to identify possible phosphorylation site motifs at this position in the protein. The site in this protein was matched to the G protein-coupled receptor kinase motif. Note that the PhosphoMotif_finder tool matched the Human site to the Casein kinase II and the G protein-coupled receptor kinase 1 motifs. Of the fish species only the spotted green pufferfish has a residue at this position, which is compatible

with phosphorylation (annotated by the cross-annotation procedure). This site was matched to the following motifs (using the same tool as before): the beta-adrenergic receptor kinase substrate, the G protein-coupled receptor kinase 1 substrate and the Casein kinase I substrate.

Residues 15(10) and 18(13) are conserved amongst all p53 homologues with a PF08563 domain annotation. All species except those which are fish have annotations for both of these positions in the Swiss-Prot database. The PTMDB contains annotations for the remaining species (all fish), these were created by the cross-annotation procedure. The surrounding residues are highly conserved hinting at the conservation of a kinase recognition site. Residue 18(13) is phosphorylated by Casein kinase I in the presence of DNA-PK (DNA dependent protein kinase) phosphorylated residue 15(10) (Dumaz, Milne, and Meek 1999).

Residue 20(15) is interesting because only the mammalian homologues have a residue that can be phosphorylated at this position. Between the mammalian sequences there are only 6 amino acid differences observed in a window size of 17 amino acids - affecting only five of the 19 mammalian species. Note that in addition to the fish p53 homologues the only amphibian represented (*Xenopus laevis*) is also missing this site. This residue has been shown to be phosphorylated by Casein kinase II, which causes the stabilisation of p53 through the loss of the E3 ubiquitin ligase Mdm2 (Hirao *et al.* 2000). Residue 20(15) has been shown to be essential for the stabilisation of p53 in response to IR and UV light (Chehab *et al.* 1999). Mdm2 is not restricted to mammalian species and has for example been identified and characterised in the African clawed frog as being able to bind to p53 homologues (Marechal *et al.* 1997). The obvious question is how p53 is stabilised (via the release of Mdm2) in species which lack Ser-20(15).

Section 6.2.3 Discussion

It is unfortunate that this analysis could only be conducted for phosphorylation sites. However it does raise a very interesting question. Why are the residues that have annotations for other modification classes not present in Pfam domains? It was previously stated that only 52% of annotation sites imported

from Swiss-Prot (excluding predicted sites) and Phospho.ELM could be overlaid onto a Pfam domain. It was also shown in Section 4.4.2 that the probability of a residue position being present in a Pfam domain decreased the closer it was to the n or c terminus of a protein. The PTMDB contains annotations for five residues in the *H. sapiens* p53 orthologue with annotations for methylation and acetylation. Four of these residues are close to the c-terminal of the protein and one is found between the DNA binding domain and the tetramerization domain. The obvious hypothesis is that these residues are located in regions that show high variability between orthologues (likely to be disordered regions).

Seven phosphorylation sites were identified in the PTMDB database for 42 p53 family members from 24 species. Two sites, at residues 99 and 215, reside in the DNA binding domain; serine 215 was present in all p53 family members. The Swiss-Prot database contains phosphorylation annotations for residue 99 in a number of p53 proteins, but not any p63 or p73 proteins. Residue 99 corresponds to a serine residue in all p63 and p73 proteins and all p53 proteins (except those from the bony fish (although in *O. mykiss* this residue is a threonine). A literature search revealed no reports of residue 99 being phosphorylated in p63 or p73 proteins. Although a report was found for the modification of residue 99 in p53 by casein kinase II, one of the enzymes predicted by the PhosphoMotif_tool to modify this residue in p63 and p73 proteins.

Five phosphorylation sites (6, 9, 15, 18, and 20) were identified in the TA domain of p53 proteins. It is interesting that not all p53 proteins have annotations for the corresponding Pfam domain. Five fish p53 proteins and one from *G. gallus* were missing corresponding annotations. In addition none of the p63 and p73 proteins had annotations for the corresponding Pfam domain. It has widely been reported that p63 and p73 proteins have TA domains. Based on multiple sequence alignments of the p53 family, it appears as though significant divergence has occurred between p53 and p63 and p73 proteins in this domain. It is therefore probable that the Pfam p53 TA domain has been trained on only p53 protein sequences. Obviously it could be argued that if it is proving difficult to identify homologous residues between the TA domains of p53 proteins and p63 and p73 proteins that a new Pfam domain should be created.

The phosphorylation of residue 20 has been shown to result in the dissociation of Mdm2, leading to the stabilisation of p53. A phosphorylation compatible residue at position 20 is only present in mammalian p53 proteins. Mdm2 homologues have been identified in other species, for instance in *X. laevis*. This raises the question of how Mdm2 is triggered to dissociate from non-mammalian p53 proteins. Phosphorylation compatible residues at positions 6 and 9 appear to be absent from all ungulate species (e.g. *O. aries*, *B. taurus*, etc.) and related species (e.g. *D. leucas*). In contrast the majority of non-ungulate mammalian species possess such residues. The exceptions being *C. lupus*, which is missing site 9, and *F. catus*, which is missing site 6. Phosphorylation compatible residues at positions 15 and 18 are seen across all p53 proteins.

Using PTM Browser an analysis into the conservation of a subset of p53 family member phosphorylation sites has been performed. The tool makes this type of analysis relatively straight forward by automatically generating multiple sequence alignments, which are starred to show modification sites. In addition the phylogenetic trees that are constructed aid in understanding the relationships between proteins. One of the main benefits of using PTM Browser is that it produces a table that documents which sites have been observed in individual proteins. The main drawback of using PTM Browser is that it is only able to compare modification annotations for residues that are located in Pfam domains. This restriction prevented many p53 modifications from being analysed. Going forward it will be very important to allow PTM Browser to analyse the conservation of modifications outside of Pfam domains. This would be a relatively straight forward process as PTM Browser already creates multiple sequence alignments of the proteins that it has been asked to analyse the conservation between. There is the obvious temptation to remove the dependency on the Pfam domain alignments entirely for protein family analysis. However it was noted during the p53 analysis that many modified residues aligned in the Pfam domains alignments, but not in the CLUSTAL W, generated multiple sequence alignments. Manual inspection of the differences suggested that, for the most-part, it was the CLUSTAL W alignment that was incorrect (although of course this is only speculative).

Section 6.3 *Taxonomic comparisons*

Section 6.3.1 Conservation between super-kingdoms

In this section the following question is asked: Are there any PTMs that are conserved across super-kingdom boundaries? Asking this question should reveal a set of PTMs that are most likely essential to the majority of Bacteria and Eukaryotes. This question can be likened to that which scientists have previously asked regarding the minimal genome that can sustain life. Indeed if some of the conserved PTM sites are missing from a minimal genome, it might suggest that a gene has been missed out.

PTM Browser has been used to identify which experimentally verified modification sites are conserved between Bacterial and Eukaryotic species (based on Pfam domain, alignment column, and PTM class). A total of 26 conserved modification positions were identified in 16 different Pfam domains – list shown in Figure 58. The majority of the domains are involved in energy metabolism, transcription and translation. The following PTM classes had at least one conserved residue in the extracted PTM annotation set: Phosphoprotein, Methylation, FMN, TPQ, Oxidation and Other.

Pfam Accession	Description
PF00016	Ribulose biphosphate carboxylase large chain, catalytic domain
PF00069	Protein kinase domain
PF00072	Response regulator receiver domain
PF00113	Enolase, C-terminal TIM barrel domain
PF00334	Nucleoside diphosphate kinase
PF00472	RF-1 domain
PF00502	Phycobilisome protein
PF00512	His Kinase A (phosphoacceptor) domain
PF00884	Sulfatase
PF00989	PAS fold
PF01116	Fructose-bisphosphate aldolase class-II
PF01179	Copper amine oxidase, enzyme domain
PF01627	Hpt domain
PF01965	DJ-1/PfpI family
PF02866	lactate/malate dehydrogenase, alpha/beta C-terminal domain
PF03143	Elongation factor Tu C-terminal domain

Figure 58: Pfam domains with experimentally verified PTM annotations conserved between Bacterial and Eukaryotic proteins

A comparison was also performed by using all annotations extracted from the Swiss-Prot and Phospho.ELM databases, thus excluding all cross-annotated sites, regardless of evidence qualifier. This comparison revealed 43 conserved modification annotations located on 28 Pfam domains. In addition to the previous list of PTM classes, this dataset also had the following: Pyruvate, Glycosylation_N_Linked, Nucleotide-binding and Organic radical. Just under half of the residues with a conserved modification were phosphorylated (see Figure 59).

PTM Class	Number of modified residues
Phosphoprotein	21
Other	8
FAD	3
Methylation	3
Glycosylation_N_Linked	2
FMN	1
Nucleotide-binding	1
Organic radical	1
Oxidation	1
Pyruvate	1
TPQ	1

Figure 59: Number of modified residues conserved between Eukaryotic and Bacterial species, grouped by PTM class. Original dataset included all PTM annotations from the Swiss-Prot and Phospho.ELM databases, regardless of evidence qualifier.

In this dataset two N-linked glycosylation sites appear to be conserved between Bacteria and Eukaryotes. Both sites fall in the Pfam Peptidase_S8 (PF00082, Subtilase, serine protease family) domain at alignment positions 1063 and 1303. An alignment of the surrounding sequences from proteins that have N-linked glycosylation annotations at these two positions is shown in Figure 60. 1063 has an N-linked glycosylation annotation for a peptidase from *Thermus sp.* (strain Rt41A) and *Magnaporthe oryzae* (Rice blast fungus). 1303 has N-linked glycosylation annotations from one Bacterial (*Thermus sp.* (strain Rt41A)) and seven Eukaryote species. All of these N-linked glycosylation site annotations have been imported from the Swiss-Prot database with the evidence qualifier **Potential**. The UniProtKB entry for the *Thermus sp.* peptidase states that the protein is a glycoprotein (Peek *et al.* 1992). The regions surrounding both sites appear to be well conserved between *Thermus sp. Rt41A* and the Eukaryote

species. However a literature search revealed no direct evidence for the glycosylation of proteins by a species of the genus *Thermus*.

A comparison was also performed between Eukaryote and Archeal PTM annotations (extracted from Swiss-Prot and Phospho.ELM). Nine conserved sites were identified from nine different Pfam domains (Figure 61). Like in the previous comparison the domains are predominantly involved with energy and nucleotide metabolism.

234 <i>Thermus sp.</i> (strain Rt41A) (1063)		
P58371	Magnaporthe grisea []	VKVLKS*NGSGT
P80146	Thermus sp. Rt41A []	VRVLDC*NGSGS
305 <i>H. sapiens</i> (1303)		
P12547	Aspergillus oryzae []	AI*NMSLG-
P28296	Aspergillus fumigatus []	AI*NMSLG-
P35211	Aspergillus flavus []	AI*NMSLG-
P80146	Thermus sp. Rt41A []	VI*NMSLG-
Q14703	Homo sapiens [human]	VL*NLSIGG
Q9WTZ2	Mus musculus [house mouse]	VL*NLSIGG
Q9WTZ3	Rattus norvegicus [Norway rat]	VL*NLSIGG
Q9Z2A8	Cricetulus griseus [Chinese hamster]	VL*NLSIGG

Figure 60: Alignment of a conserved N-linked glycosylation site in the Pfam domain PF00082.

Alignment column 1063 corresponds to position 254 in P58371 and 234 in P80146

Pfam Accession	Description
PF00016	Ribulose biphosphate carboxylase large chain, catalytic domain
PF00072	Response regulator receiver domain
PF00113	Enolase, C-terminal TIM barrel domain
PF00180	Isocitrate/isopropylmalate dehydrogenase
PF00221	Phenylalanine and histidine ammonia-lyase
PF00467	KOW motif
PF00543	Nitrogen regulatory protein P-II
PF01979	Amidohydrolase family
PF03764	Elongation factor G, domain IV

Figure 61: Pfam domains with PTM annotations conserved between Eukaryotic and Archeal species.

Includes only those PTM annotations imported from the Swiss-Prot and Phospho.ELM databases.

A three-way-comparison was also performed between the three superkingdoms. This comparison identified six modification sites that were found in all three superkingdoms (shown in Figure 62).

Pfam Accession	Alignment Position	PTM Class
PF00072	317	Phosphoprotein
PF00113	333	Phosphoprotein
PF00016	67	Other
PF00221	467	Other
PF01979	747	Other
PF00543	100	Nucleotide-binding

Figure 62: Modification sites conserved across all three superkingdoms
Includes only those modifications extracted from the Swiss-Prot and Phospho.ELM databases.

Section 6.3.2 Conservation between Eukaryote species

In this section the following question is asked: To what extent are post-translationally modified proteins conserved between different Eukaryote species? The answer to this question could, for instance, suggest that a particular species is a poor model of the *H. sapiens* phosphoproteome.

The first part of this section looks at the conservation of proteins between eight species and *H. sapiens*, based on five different modification classes. The second part of this section looks in more detail at the conservation of modifications between *H. sapiens* and *M. musculus*.

Section 6.3.2(a) Conservation of modified proteins between Eukaryote species

To keep the number of comparisons being discussed to a manageable level all species have been compared to *H. sapiens*. Eukaryote species were first selected whose non-redundant protein sets (for simplicity referred to as proteomes), as created in Chapter 3, approximately matched the number of protein coding genes. The following seven species had almost a 1:1 ratio between the number of proteins and protein coding genes: *M. musculus*, *T. nigroviridis*, *D. melanogaster*, *C. elegans*, *N. vectensis*, *A. thaliana*, *S. cerevisiae* and *S. pombe*. Note that the estimated number of protein coding genes for *T. nigroviridis* has previously been discussed as being a gross underestimate - this species will therefore also be looked at in this section. A brief analysis of the conservation of the proteins from these species and *H. sapiens* was previously presented in Table 27 - these numbers can also be found in Table 42 along with the estimated proteome completeness. PTM Browser was used to calculate the conservation of proteins with different PTM

class annotations between *H. sapiens* and each of the other eight Eukaryote species. To accomplish this PTM Browser makes use of the orthologue and in-paralogue assignments added to the PTMDB using the CoPaO software (see Chapter 3). Only results for the following five PTM classes are discussed here: Phosphorylation, N-linked Glycosylation, O-linked Glycosylation, GPI anchor, and Acetylation. Results from this analysis are shown in Table 42.

The most obvious result is that Phosphorylated and Acetylated proteins appear to be more conserved than the proteomes as a whole between all eight species and *H. sapiens*. Another interesting result is that even after the cross-annotation process neither *S. cerevisiae* nor *S. pombe* have any proteins with a GPI-anchor annotation. It's interesting to note that *H. sapiens* has 64 proteins with a GPI-anchor, the only species with similar numbers were *M. musculus* (74) and *T. nigroviridis* (40).

The number of proteins with O-linked Glycosylation sites seems to dramatically fall as the evolutionary distance to *H. sapiens* increases. For example, 1,104 *H. sapiens* proteins have O-linked Glycosylation annotations, which is similar to the number for *M. musculus* (1,005). *T. nigroviridis* had 633, which falls to 188 for *A. thaliana* and finally falls to only 11 and 9 for *S. cerevisiae* and *S. pombe*, respectively. For all but two species (*M. musculus* and *T. nigroviridis*) N-linked glycosylated proteins are conserved to a greater extent than O-linked. It does at first sight appear rather strange that proteins predominately involved in cell-cell signalling are more highly conserved in single celled organisms, compared to O-linked proteins whose functions almost exclusively exist within the cell. Further work will need to be done to understand why N-linked glycosylated proteins show a higher degree of conservation. For the majority of PTM classes a larger percentage of the other species' modified proteins are homologous to proteins from *H. sapiens*, rather than the other way around. There are some exceptions, for example 51.5% of *T. nigroviridis* phospho-proteins have *H. sapiens* orthologues compared to 61.3% in the reverse orientation.

	Proteome Completeness	Proteome conservation	Phosphorylation	N-linked Glycosylation	O-linked Glycosylation	GPI-anchor	Acetylation
<i>Mus musculus</i>	104.63%	72.57%	80.30%	63.95%	78.71%	61.33%	88.23%
<i>Homo sapiens</i>		62.48%	77.11%	65.42%	73.73%	67.19%	82.33%
<i>Tetraodon nigroviridis</i>	172.55%	39.47%	51.50%	38.38%	44.55%	20.00%	54.58%
<i>Homo sapiens</i>		45.97%	61.34%	49.75%	55.89%	25.00%	61.43%
<i>Drosophila melanogaster</i>	108.72%	39.85%	56.24%	34.34%	34.63%	0.00%	67.98%
<i>Homo sapiens</i>		31.20%	42.82%	25.29%	20.02%	14.06%	56.76%
<i>Caenorhabditis elegans</i>	99.23%	25.33%	46.31%	27.08%	27.73%	16.67%	44.33%
<i>Homo sapiens</i>		28.82%	41.55%	29.72%	17.03%	12.50%	53.74%
<i>Nematostella vectensis</i>	100.03%	31.50%	42.44%	23.96%	27.01%	14.29%	58.62%
<i>Homo sapiens</i>		33.74%	43.88%	35.13%	21.29%	15.63%	55.61%
<i>Arabidopsis thaliana</i>	104.80%	27.02%	48.50%	28.17%	43.62%	0.00%	56.93%
<i>Homo sapiens</i>		19.33%	26.34%	10.73%	4.08%	17.19%	45.11%
<i>Saccharomyces cerevisiae</i>	99.91%	32.72%	51.22%	29.37%	27.27%		64.91%
<i>Homo sapiens</i>		13.57%	19.72%	5.89%	0.36%	0.00%	30.56%
<i>Schizosaccharomyces pombe</i>	99.14%	45.29%	58.62%	27.10%	77.78%		66.98%
<i>Homo sapiens</i>		12.71%	16.82%	6.04%	0.45%	0.00%	30.04%

Table 42: Conservation of modified proteins between *H. sapiens* and a selection of other Eukaryote species .

Proteome completeness, expresses the correlation between the number of proteins in the non-redundant protein set created from UniProtKB release 13.3 (as extracted using the protocol in Section 3.4.3) and the number of protein coding genes listed for an organism in the Ensembl database (and other resources); values extracted from Table 23. Proteome conservation, represents the percentage of an organisms proteome that has an orthologous partner in another organisms proteome. The methodology used to incorporate orthologue annotations into the PTMDB has already been described in Chapter 3. The proteome conservation values shown in this table have been extracted from Table 27; note that in-paralogues have been included in the underlying counts (in-paralogues are by definition orthologues). The remaining columns show the conservation of proteins subdivided by PTM class; these values were obtained from PTM Browser. Note that the following PTM annotations were included in the PTM Browser analysis; Swiss-Prot [Probable, Experimental, and By similarity], Phospho.ELM [LTP and HTP], and all cross-annotations that were homologous to Eukaryote sites from Swiss-Prot [excluding predicted sites] or Phospho.ELM.

Section 6.3.2(b) *H. sapiens* and *M. musculus*

H. sapiens and *M. musculus* are the only mammals in the PTMDB whose proteome size matches or exceeds their estimated number of protein coding genes (see Section 3.4.3(f)). The *H. sapiens* proteome was previously estimated to be 121% complete and the *M. musculus* 105%. Proteomes in the PTMDB may exceed the estimated number of protein coding genes where splice variants are counted, or the number of protein coding genes is an underestimate.

A conservation analysis was first performed at the domain level – the aim of which was to identify conserved modifications based on the triplet Pfam accession, Pfam alignment position, and PTM class. It must be stressed that this analysis does not involve the direct comparison of orthologues, as identified by the CoPaO software. The percentage of conserved domain level sites was calculated for each different PTM class. An annotation was considered to be conserved if the corresponding values for the fields Pfam accession, Pfam alignment position, and PTM class could be found together in the dataset for the other species. This analysis was first carried out using only those annotations that had been imported from Swiss-Prot and Phospho.ELM – the **core set**. The analysis was then carried out by adding annotations from the cross-annotation set to those used in the first analysis – the **main set**. Note that cross-annotations for sites that didn't have a homologue in either the *H. sapiens* or *M. musculus* Swiss-Prot or Phospho.ELM datasets were excluded.

The percentage of conservation observed for a number of particularly important PTM classes is shown in Table 43. The final column of this table shows the degree of conservation that was observed for the core set. Palmitoylation, myristoylation, and prenylation are all lipid modifications that are important in the localisation of proteins to plasma membranes. Almost all of the modification sites for these three PTM classes are conserved between *H. sapiens* and *M. musculus*. 32% of the *H. sapiens* phosphorylation sites do not have a corresponding site in the *M. musculus* core dataset. After the cross-annotation process only 7% of the *H. sapiens* phosphorylation sites do not have a homologue in *M. musculus*. The core dataset appears to show that only 43% of the *H. sapiens* O-linked glycosylation sites are conserved with *M. musculus* –

after cross-annotation this number jumps to 87%. This analysis clearly showed an exceptional level of conservation for all PTM classes – that for many approaches 100%.

PTM Class	Total	Main set		Core
		Intersect	%	%
<i>Phosphoprotein-H. sapiens</i>	3618	3373	93.23	68.31
<i>Phosphoprotein-M. musculus</i>	3500		96.37	87.93
<i>Glycosylation_N_Linked-H. sapiens</i>	4692	4314	91.94	70.43
<i>Glycosylation_N_Linked-M. musculus</i>	4644		92.89	76.78
<i>Glycosylation_O_Linked-H. sapiens</i>	207	179	86.47	42.64
<i>Glycosylation_O_Linked-M. musculus</i>	193		92.75	77.78
<i>Acetylation-H. sapiens</i>	267	260	97.38	93.77
<i>Acetylation-M. musculus</i>	295		88.14	84.27
<i>Methylation-H. sapiens</i>	41	40	97.56	94.87
<i>Methylation-M. musculus</i>	41		97.56	92.50
<i>Myristate-H. sapiens</i>	11	11	100.00	63.64
<i>Myristate-M. musculus</i>	11		100.00	100.00
<i>Palmitate-H. sapiens</i>	52	52	100.00	100.00
<i>Palmitate-M. musculus</i>	52		100.00	98.08
<i>Prenylation-H. sapiens</i>	3	3	100.00	100.00
<i>Prenylation-M. musculus</i>	3		100.00	100.00

Table 43: Conservation of modified domain residues between *H. sapiens* and *M. musculus*.

Swiss-Prot and Phospho.ELM PTM annotations were extracted from the PTMDB for both species – forming the core set. In addition PTM annotations were extracted from the cross-annotation section of the PTMDB for both species. Cross-annotations were limited to those that had a matching PTM annotation in the core set. The core set and the restricted cross-annotation set form the main set.

Some of the cross-annotations will of course be false positives, as will some of the PTM annotations in the core dataset (i.e. those imported from Swiss-Prot with the evidence qualifier: Potential). The true degree of conservation between these two species most likely lies somewhere between that observed for the core set and that of the main set. The main conclusion from this analysis appears to be that differences in the modification of residues in conserved domains are more likely to be caused by upstream changes in the enzymes responsible for carrying out the modification – rather than the absence of compatible residues.

Section 6.3.3 Discussion

This section started with an analysis into the conservation of modifications between super-kingdoms. The first comparison used only PTM annotations with an experimental evidence qualifier. Even though this first dataset was

extremely small (compared to the total number of annotations in the PTMDB) 26 modification sites were conserved between Eukaryote and Bacterial species. The fact that a small number of modification positions are conserved may be surprising, however what is perhaps not surprising is that these sites located to protein domains, involved in energy metabolism, transcription and translational (core cellular processes). The second comparison involved all PTM annotations imported from the Swiss-Prot and Phospho.ELM databases (no longer just the experimentally verified sites - although cross-annotations were still excluded). This analysis expanded the number of conserved modification sites to 43 belonging to 28 Pfam domains. Perhaps the most interesting conserved modification was that of an N-linked glycosylation site, that appears to be conserved between a Eukaryote and Bacterial protein. The conserved modification was observed between the Bacterial species, *Thermus sp.* (strain Rt41A), and the Eukaryote, *Magnaporthe oryzae*. The modification site was located in the Subtilase; serine protease family domain. It must be emphasised that some Bacteria can carry out N-linked glycosylation. Obviously the conservation data presented in this section, does not preclude the process of convergent evolution. Note that only six modification sites were conserved across all three super-kingdoms.

The second analysis presented, looked at the conservation of proteins grouped by PTM class. Specifically this section looked at the percentage of *H. sapiens* proteins with particular modification classes that had orthologues in a range of other Eukaryote species. The results of this analysis are presented in Table 42. It is worth noting that the *H. sapiens* proteome used in this study was estimated to be 121% complete. It is unclear just how much of the 21% over-estimate is related to redundancy/erroneous sequences in the proteome used or an under-estimate in the number of protein coding genes (see Section 3.6 for a more detailed analysis of this issue). However, what is clear is that if a significant portion of the 21% represents erroneous sequences (i.e. fragments); this could have a profound effect on the results presented in this section. The reason is simply that fragmentary sequences are unlikely to be assigned to an orthologue group (and nor should they), which will in turn over-estimate the number of proteins in the *H. sapiens* proteome that do not have orthologues in other

species. In addition many of these erroneous sequences are likely to be missing Pfam domain annotations (and thus less likely to have PTM annotations in the PTMDB (because they were excluded from the cross-annotation process)). When both of these issues occur, the result would most likely manifest as the percentage of proteins conserved between the *H. sapiens* proteome and another species being lower than that seen for PTM specific groups of proteins. This issue will not be discussed any further, but readers are encouraged to bear this point in mind.

Phosphorylated and acetylated proteins appeared to show the greatest degree of conservation; which is not surprising given the fact that many phosphorylation and acetylation sites have been shown to occur on key cellular regulators (i.e. p53). GPI-anchors are associated with proteins that are involved in cell signalling, it should therefore perhaps not come as a complete surprise that neither *S. pombe* nor *S. cerevisiae* contain any proteins that are orthologous to the *H. sapiens* proteins with GPI-anchor annotations. It is interesting that this modification seems to be more prevalent in species that are more closely related to *H. sapiens*. Although this was true of all PTM classes analysed in this section, it is just more aberrant for GPI-anchored proteins because there are so few of them. Like GPI-anchored proteins, many species, have few orthologues to the *H. sapiens* O-linked glycosylated proteins. These results suggest that not only are some PTM classes fairly recent evolutionary inventions but that they are carried by proteins that are not even present in species that cannot carry out such modifications.

The final analysis looked at the conservation of modifications between *H. sapiens* and *M. musculus*. Note that as this analysis was being carried out at the domain level it was unaffected by any potential issues with the *H. sapiens* proteome being larger than expected. For each modification class the conservation was analysed both before and after the inclusion of cross-annotation results. Cross-annotations were excluded if a homologous annotation could not be found in the PTM annotations imported from the Swiss-Prot or Phospho.ELM databases, on *H. sapiens* or *M. musculus* proteins. For the majority of PTM classes analysed almost all of the modification sites were conserved between *H. sapiens* and *M. musculus* proteins in the original

annotations imported from the Swiss-Prot and Phospho.ELM databases. One notable exception was that only 68.31% of the *H. sapiens* phosphorylated sites were present in the *M. musculus* PTM annotations imported into the PTMDB. After cross-annotation this number jumped up to 93.23%; of course not all of the cross-annotations will be correct. Another interesting difference was that only 42.68% of the *H. sapiens* O-linked glycosylation sites were conserved in the *M. musculus* PTM annotations imported into the PTMDB. After cross-annotation this percentage rose to 86.47%. The main message from these results appears to be that the vast majority of modification sites are shared between *H. sapiens* and *M. musculus*. What is possibly more interesting however is that the cross-annotation procedure has identified sites that are definitely not shared between *H. sapiens* and *M. musculus*. These non-conserved sites should be of interest to scientists that have chosen *M. musculus* as a model organism to study particular *H. sapiens* diseases.

Chapter 7

General Discussion

Section 7.1 *PTM Databases*

A selection of PTM annotation databases were reviewed in Section 1.2. The PTM community has created a number of very useful databases. Some of which are specific to particular PTM classes (e.g. Phospho.ELM), others contain annotations for a much wider range of PTM classes (e.g. UniProtKB). In Chapter 2 the construction of a new database of PTMs, the PTMDB, was discussed, that includes annotations from UniProtKB, Phospho.ELM and the PDB. The PTMDB contains 186,408 modifications extracted from Swiss-Prot for 56,074 proteins. The Phospho.ELM database contributed 14,934 phosphorylation annotations for 4,917 Swiss-Prot entries. In addition 656 phosphorylation annotations were extracted from Phospho.ELM for 292 TrEMBL entries.

Collating PTM annotations from pre-existing PTM databases is not always a straight forward process. PTM annotations are made available in various different formats. Each format requires a specific import procedure. Although PTM annotation formats differ, many are now adopting either the Swiss-Prot PTM controlled vocabulary or PSI-MOD ontology, to describe chemical modifications. During the creation of the PTMDB it became obvious that some databases made it difficult to unambiguously identify which residues were being referred to. When annotating UniProtKB sequences it is essential for the annotation to include not only the UniProtKB accession, but also the sequence revision number. An alternative to providing the sequence revision number is to provide the protein sequence.

There are far more observed glycan structures (e.g. KEGG/Glycan contains 11,000 structures) than all of the other PTM types put together. Based on the number of predicted glycosylation sites, it should be one of the most abundant

PTM classes. However determining both the position and structure of a glycan is a difficult process (the field of Glycoproteomics). It was shown in Table 10 that only 5% of Swiss-Prot glycosylation annotations were experimentally derived; compared to 41% of phosphorylation annotations. Of course it could be argued that many of the predicted glycosylation annotation sites in Swiss-Prot are not actually modified at all. Almost all glycobiology papers begin by stating that glycosylation is the most abundant modification. This is a somewhat bold statement given that most authors backup this claim by referring to the number of glycosylation site annotations in the Swiss-Prot database (of which >90% are predicted).

Attempts have been made to make what little information is available on site specific glycosylation available in databases (e.g. GlycoSuiteDB (Cooper *et al.* 2001) and PDB2LINUCS (Lutteke, Frank, and von der Lieth 2004)). The Swiss-Prot database currently annotates glycosylation sites with only the terminal-reducing sugar. Swiss-Prot does indirectly store a more complete record of glycan structures for a handful of glycosylation site annotations that have been cross-referenced to the GlycoSuiteDB.

In Section 2.6 the incorporation of glycosylation sites and corresponding structures extracted from the PDB using the PDB2LINUCS tool was discussed. There were three points raised by this work. The first was that the PDB did contain glycosylation sites that were not present in the Swiss-Prot database; although there were only 105. The second point is that although 1,363 residue/glycan pairs were identified, 80% of these sites were for N-linked glycans with only the first two GlcNAc residues present. The final point raised was the importance of expanding the current PTM vocabularies so that they better support glycosylation PTM types. Section 2.6.4(c) discussed how the PTMDB vocabulary was extended to support the data extracted from the PDB using the PDB2LINUCS tool. PTM vocabularies should be expanded to: a) support raw glycan structure formats (e.g. GlycoCT) and b) to support some kind of glycan structure classification (e.g. High mannose, hybrid, complex, etc.). It should also be noted that most glycosylation-specific databases also allow users to search databases by actually drawing a glycan structure (see

<<http://www.glycome-db.org/database/searchSubStructure.action>> for an example).

Section 7.2 *Cross-annotation*

In Section 2.5.1 it was pointed out that the Swiss-Prot curators transfer experimentally verified PTM annotations between homologous proteins. It was noted that this process is only carried out between closely related species (where closely is never defined explicitly). It had already been decided that the conservation analysis that was going to be carried out would be based solely on the annotations in the PTMDB. This raised an important question regarding how complete the cross-annotated dataset in Swiss-Prot actually was. A brief analysis of *H. sapiens* and *M. musculus* PTM annotations suggested that some cross-annotations were missing (from highly homologous sequence regions). Obviously some of the cross-annotations may have been missing because they were deemed inappropriate. As the Swiss-Prot annotation process is at least in-part manual, it is also possible that a curator has simply not had a chance to transfer these annotations. It was the discovery of these missing cross-annotations that first led to the decision to carry out a new cross-annotation process.

The next consideration was that it would be useful to include TrEMBL proteins in PTM conservation analysis; especially as many species are represented almost entirely in the UniProtKB by TrEMBL entries. It is important to reiterate that the TrEMBL database does not include any PTM annotations. In addition it had been decided that it would be useful to analyse PTM conservation between both distantly and closely related species. To account for all of the issues just raised the cross-annotation process was designed to transfer all annotations between all proteins in the UniProtKB. This procedure included both Swiss-Prot and TrEMBL proteins and did not restrict the transfer of annotations to that between only closely related species. The Swiss-Prot cross-annotation process is designed to produce a high confidence dataset. In contrast the PTMDB cross-annotation process was designed to produce a dataset that could be constrained later, as required. For example using the PTM Browser tool (discussed in Chapter 5) it is possible to create cross-annotation datasets that

impose the same type of constraints as used in the Swiss-Prot cross-annotation process. For example, it is possible to exclude cross-annotations that have been made between Bacterial and Eukaryote proteins.

The cross-annotation procedure was designed to transfer annotations based on the homology identified in PfamA domain alignments. Pfam domain alignments were used as these are guided by a manually curated seed alignment; which was considered to be more reliable than those generated automatically by programs such as CLUSTAL W. In addition by using the Pfam domain alignments the cross-annotation procedure did not need to include a lengthy sequence alignment step. The procedure also calculated the sequence identity between acceptor and donor sites in sequences of various window sizes. PTM Browser can be used to also constrain cross-annotations by both window size and identity count.

The obvious criticism of the cross-annotation dataset is that no attempt was made to confirm that the new PTM annotation sites conformed to known enzyme acceptor sites. Indeed during the analysis of the p53 family of proteins (see Section 6.2) this process had to be carried out manually. It should be pointed out that traditional prediction programs make no use of homology information to weight predictions higher if a modification has previously been observed at a homologous residue in another protein. Therefore in the future it should prove beneficial to combine homology based cross-annotation techniques with traditional prediction tool chains.

As the cross-annotation procedure was essentially only constrained by the Pfam domain alignments it should come as no surprise that it generated a very large number of cross-annotations. Even when cross-annotations made between super-kingdoms were excluded there were still over 2 million. It is clear from the work shown in Table 43 that the cross-annotation dataset significantly reduces the percentage of modifications that do not appear to be conserved (as always with the caveat that some of the cross-annotated sites may not be modified in real life) between some species.

The version of the Swiss-Prot database that was used in the PTMDB was frozen at the start of development as one released in 2008. At the end of this

study a newer version of the PTMDB was created that contained PTM annotations from the December 2010 release of Swiss-Prot. It is interesting to note that 3,159 of the cross-annotations are present in the newer version of Swiss-Prot.

The PTMDB cross-annotation dataset should prove to be of interest to a wide audience. Especially for scientists who work with species whose proteome is predominantly or completely represented by TrEMBL entries. For instance the PTMDB contains the first large scale PTM annotation (of which the author is aware) for the two fish *D. rerio* and *T. nigroviridis*. Another benefit of the PTMDB cross-annotation dataset is that it provides the user with everything they need to define their own constraints on what cross-annotations to allow (obviously many may argue that this is also a drawback). There is another benefit of carrying out the cross-annotation process that may not be immediately obvious. Many users might be dissuaded from using the cross-annotation dataset in its current form as it does not attempt to validate cross-annotated acceptor sites with known enzyme acceptor sites. These users might want to remember that the lack of a cross-annotation in the dataset tells them immediately that a site is almost certainly not conserved (i.e. the residue at the homologous position is not compatible with the PTM type at all). The only caveat to this point is that it assumes that its Pfam domain annotation set is complete (i.e. that domain boundaries are correct and that a domain annotation is not simply missing). This issue was previously discussed in Section 6.2.2 where almost all of the p53 proteins had Pfam TA domain annotations but neither the p63 nor p73 proteins did (although as discussed this may be deliberate).

Section 7.3 *Proteomes and orthology detection*

Chapter 3 was completely devoted to the construction of non-redundant protein sets and the detection of orthologous proteins in the PTMDB. It is worth restating that throughout this thesis the terms proteome and non-redundant protein set have been used interchangeably. Hopefully it is somewhat obvious that a non-redundant protein set is not by definition a proteome; although a proteome has to be a non-redundant protein set. The procedure outlined in

Chapter 3 creates non-redundant protein sets that are for the most part devoid of splice variant sequences (and thus for Eukaryote species they are not technically proteomes).

One of the main issues found with the proteome construction technique described in Chapter 3 is that it resulted in proteomes that exceeded the number of protein coding genes for a number of species. For some species such as *H. sapiens* it appeared as though a number of redundant Swiss-Prot sequences had been included in the proteome (remember that Swiss-Prot entries are meant to be non-redundant). For others, such as *T. nigroviridis*, it appears as though the estimated number of protein coding genes was incorrect.

In order to detect orthologues in the PTMDB the InParanoid technique was adopted. Although we had obtained the InParanoid software from the original authors to run across our database – it failed to produce any orthologue assignments. The decision was therefore taken to re-implement the InParanoid algorithm as a new Perl program called CoPaO. Initial results comparing *H. sapiens* and *M. musculus* suggested that the new Perl program would benefit from being multi-threaded. A new Java version of CoPaO has since been created, which shows little benefit in the application of threading to the InParanoid algorithm.

During the development of PTM Browser the only other PTM resource that combined orthologue assignments with PTM annotations was PHOSIDA. At the start of 2011 new InParanoid orthologue annotations seem to have been added to the UniProtKB. It will therefore be important to monitor the inclusion of these annotations into the UniProtKB. This will most likely allow for future versions of the PTMDB to include InParanoid annotations directly from the UniProtKB. This will in turn free up resources so that additional species pairs can have their orthologues identified using the CoPaO software.

Section 7.4 *PTM Conservation analysis tools*

PTM Browser is novel in that it has been specifically designed to allow for the conservation of modifications to be analysed. PHOSIDA (Gnad *et al.* 2007; Gnad, Gunawardena, and Mann 2011) is a predominately phosphorylation

related resource that does however share some of the same functionality as PTM Browser. For a specific modification site PHOSIDA can show users which homologues (identified as being in the same IPI (International Protein Index) group) share the modification. PHOSIDA can only be used to analyse the conservation of modification sites between nine species. The number of species that are available for comparison with PTM Browser depends on the query being performed. When a user is comparing modifications between orthologues, as defined by the CoPaO software (refer to Section 3.3), they are limited to 35 species. However when they are comparing modifications based only on Pfam domain homology they have access to 7,088 species that were included in the cross-annotation process (some 2,768,148 proteins were included in this process). PHOSIDA contains 80,000 mass spectrometry derived Acetylation, Phosphorylation, and N-linked Glycosylation annotations (Gnad, Gunawardena, and Mann 2011). The majority of PTM Browser queries are likely to be limited to those sites in Pfam domains. When this constraint is in effect PTM Browser has access to 10,092 experimentally verified modification sites (26,332 when Swiss-Prot by similarity annotations are included). Obviously in addition to the experimentally derived annotations PTM Browser can also include annotations from the new cross-annotation process and those marked as Potential in the Swiss-Prot database. PHOSIDA has been designed to show users the conservation of individual modifications. In contrast PTM Browser shows users which modification sites are present in a group of proteins and most importantly which proteins have which modification sites. PHOSIDA can only display the conservation of a modification between members of the same IPI group. PTM Browser can be given a list of proteins for which modification conservation should be shown. It is worth noting that the p53 family analysis described in Section 6.2 would have been much more complex using PHOSIDA as it lacks this functionality. PTM Browser can also be asked to show the conservation for the orthologues of a particular set of proteins (using the CoPaO dataset).

PTM Browser is the first resource that enables users to identify both conserved and non-conserved modifications at the taxonomic level. In Section 6.3 such an analysis was performed to identify which modification sites were shared

between each of the three super-kingdoms. PTM Browser can also be used to calculate the percentage of proteins with particular modifications that have orthologues in other species. Such an analysis is presented in Section 6.3.2(a).

In conclusion PTM Browser presents users with new workflows (see Section 5.3) that should dramatically reduce the amount of time taken to analyse PTM conservation.

Section 7.5 *Conservation analysis conclusions*

In Section 6.2 a review was presented on the conservation of PTMs between members of the p53 tumour suppressor family. The analysis provides an example of using PTM Browser to compare the conservation of PTMs both across species (i.e. *H.sapiens* p53 compared to *M.musculus*) and between paralogues (i.e. *H.sapiens* p53 compared to the related p63 protein). p53 has been documented as being subjected to at least six different PTM processes (i.e. phosphorylation, methylation, acetylation, ubiquitination, sumoylation, and neddylation). One of the first conclusions from this analysis was that only the phosphorylation sites were located in conserved protein domains (Pfam A domains). A brief analysis of the other sites appeared to show that they were located on the periphery of the conserved protein domains. Further work could characterise the regions in which the other sites are located (i.e. substitution rate, secondary structure, etc.). Of course a PTM conservation analysis that is only able to include sites found in conserved protein domains is incomplete. In the future it will be important to develop PTM Browser so that it can analyse the conservation of PTMs found at sites not annotated to conserved protein domains. The most obvious methodology to implement this feature would be to deduce homologous sites between orthologues using a multiple sequence alignment tool such as CLUSTAL W (Thompson, Higgins, and Gibson 1994).

The analysis (see Section 6.2.1) showed that the p53 DNA binding domain doesn't appear to have gained new phosphorylation sites over evolutionary time. The PTMDB contains only two phosphorylation sites for this domain. These two sites are present in all proteins identified as p63 and p73. These two phosphorylation sites are also found in all copies of the p53 protein except those belonging to the bony fish. Considering that it is thought that p53 and a

p63/p73-like gene (which was subsequently duplicated to form p63 and p73 genes) were the products of a single gene duplication event (Belyi *et al.* 2010), this result raises further questions. This result suggests that sometime after the divergence of the ancestor of mammals and amphibians from that of the extant bony fish, the latter lost one of the two phosphorylation sites. This analysis showed that the phosphorylation site located at residue 215 on the *H.sapiens* p53 orthologue can be found in all members of the p53 protein family. This residue has been shown to be important in preventing apoptosis after DNA damage (Liu *et al.* 2004).

The analysis of the p53 TA domain presented a more complex pattern of PTM conservation than that of the p53 DNA binding domain. A comparison between all three members of the p53 family was not possible as Pfam lacks corresponding domain annotations for p63 as well as p73 orthologues. In addition five of the bony fish p53 orthologues didn't have corresponding annotations. It was previously speculated that poor sequence conservation is likely to be the cause of these proteins not receiving TA domain annotations (see Section 6.2.2). Of course if an analysis routine were to be implemented that could utilise CLUSTAL W alignments in addition to Pfam domains these proteins could have been included in the analysis. Although the CLUSTAL W alignments would need to be treated with caution where sequence identity is low.

The PTMDB contains five phosphorylation annotations for the p53 TA domain. Three of these are conserved across all mammalian species analysed. The remaining two were found in all of the primates and other non-ungulate species (with *F. catus* and *C. lupus* being exceptions – where they each are missing one of these two sites). It is therefore possible to speculate that the non-ungulate ancestor evolved these two sites not seen in the corresponding ungulate ancestor. Although *Sus. Scrofa* (an ungulate) and *X. laevis* (an amphibian) violate this model of evolution as they both appear to possess one of these two sites. Of course more complex models of evolution could be suggested that attempt to explain these observations (i.e. convergent evolution and/or loss of sites from the ungulate ancestor). However it must be stressed that since many of these phosphorylation sites have been cross-annotated many may not

actually be modified *in vivo*. Indeed many of the sites which received a cross-annotation showed a markedly different kinase prediction profile (i.e. the enzymes that were predicted to be able to modify such sites) to that of the donor site. This of course highlights one of the issues of analysing PTM conservation with a dataset that contains predicted PTM annotations. The analysis of this domain raised an interesting question regarding the mechanism that causes MDM2 to disassociate from p53 (see Section 6.2.2). Phosphorylation of residue 20 on the *H. sapiens* p53 orthologue has been shown to be essential for the dissociation of MDM2. However *X. laevis* is missing this phosphorylation site even though it has been shown to have a MDM2 orthologue.

In Section 6.3.1 an analysis of the conservation of modifications between the super-kingdoms showed only a few modifications to be conserved. Macek *et al.* (2008) has previously looked at the conservation of phosphorylation sites between Eubacteria and Archaea. They also reported that some phosphorylation sites are conserved between Archaea and Eukaryotes. The study in Section 6.3.1 described two phosphorylation sites that are potentially conserved across all three superkingdoms.

Macek *et al.* (2008) also looked at the conservation of phosphoproteins between *E. coli* and other bacterial species. They discovered that phosphoproteins are more highly conserved than their non-phosphorylated counterparts. They found that >50% of *E. coli* phosphoproteins were conserved compared to 25% of non-phosphorylated proteins. In Section 6.3.2(a) a similar analysis was performed between Eukaryote species. This analysis showed a similar pattern of modified proteins of various classes showing a higher degree of conservation than their non-modified counterparts.

In Section 6.3.2(a) it was shown that species which had no PTM annotations for a particular class also lacked orthologues for proteins from other species that have annotations of the corresponding class. For example *S. cerevisiae* and *S. pombe* have no orthologues to *H. sapiens* proteins that have GPI-anchor modification annotations. Therefore the data presented here indicates that the

evolution of modification processes is connected to the evolution of the proteins which are modified. This is a novel finding that requires further validation.

Section 7.6 *Biological relevance*

PTM Browser can be used during the design of new drugs to identify suitable animal models. Consider the following case-study. TBN1412 is a monoclonal antibody that was designed by TeGenero to activate regulator T-cells which it was hoped would aid in the treatment of autoimmune diseases and cancer (Attarwala 2010). TBN1412 was designed to activate regulator T-cells by interacting with and activating signalling downstream of CD28 (a cell surface glycoprotein) (Attarwala 2010). Before phase I clinical trials this drug was primarily tested using macaques as the animal model (Bhogal and Combes 2007). Macaques were chosen according to TeGenero as the region of CD28 they were targeting was 100% identical to that of humans (Bhogal and Combes 2007). All volunteers that received the drug had an immediate adverse reaction to the drug that resulted in multi organ failure (Attarwala 2010). A subsequent investigation into the drug trial reported that neither contamination nor mismanagement of the trial could be blamed for the unexpected side-effects (Attarwala 2010). A number of theories have since been put forward to explain why the animal models failed. One suggestion is that the drug resulted in a different set of downstream signalling events in humans than in the macaques (Bhogal and Combes 2007). CD28 contains a number of putative glycosylation sites and has been determined to be approximately 50% carbohydrate by mass (Ma *et al.* 2004). Ma *et al.* (2004) previously showed that CD28 is negatively regulated by N-linked glycans. One possible explanation is that TBN1412 was more responsive in humans than macaques because of differences in attached glycans. Bhogal and Combes (2007) reported that Macaque CD28 wasn't in fact 100% identical to that of humans: as previously reported by TeGenero. However they also reported that all of the putative N-linked glycosylation sites were conserved; this doesn't rule out a difference in glycosylation site occupancy (i.e. which sites have which glycans). This case-study shows how important it is to pay close attention to differences between orthologues. It also provides a speculative example of where differences in glycosylation between orthologues may have resulted in the failure of a potential drug therapy. PTM

Browser presents scientists with a new tool to determine if modification sites are conserved between orthologues which could help in the selection of animal models, in future drug trials.

Section 7.7 *Future direction*

New PTM databases appear on a regular basis and it will be important to make sure that these new developments are incorporated into the PTMDB. For example a new Glycosylation database, called UniCarb-DB, was recently released (Hayes *et al.* 2011). In addition PTM Browser doesn't have access to the PDB derived residue/glycan structure pairs incorporated into the PDB. They were excluded partly because of their small number and partly because they add a layer of complexity to the display format of PTM Browser. Full support for such datasets should be added to PTM Browser as it seems likely that over time the size of these datasets will expand dramatically. PTM Browser will need support for displaying glycan structures and for users inputting specific structures.

PTM Browser was originally designed to only be able to compare annotations based on PfamA domain alignment coordinates. There is a definite need to be able to analyse PTM conservation outside of PfamA domains and outside of Pfam as a whole (only 50% of modifications fall inside a PfamA domain). This issue could partly be addressed by adding support for PfamB (see Section 4.3.1(a) for more details on the exclusion of PfamB domains). However as was evident from the analysis of the p53 family of proteins modifications can lay outside of PfamA and PfamB domains. To allow PTM Browser to better accommodate protein family analysis, it should be extended so that for modifications that lay outside of Pfam domains, their conservation is derived from new multiple sequence alignments (e.g. use the automatically generated CLUSTAL W alignments that PTM Browser already produces). However it should be stressed that where available the Pfam domain alignments should be used preferentially - as these are guided by manually curated seed alignments.

PTM Browser is a powerful application in terms of what modifications users can select to compare and how the comparison is actually performed. Going forward it will be important to develop the PTM Browser GUI to make it much

easier for users to access the common types of queries. In addition to making changes to the web application, it will be important to make the developer web service as easy to use as possible. To this end a series of wrapper classes are already under development that hide a lot of the complexity of the web service from developers (e.g. automatically dumping datasets to a database). There are many other features that could be added to the PTM Browser interface. For example one idea would be to generate a tree like structure similar to that created by PHOG. Instead of showing which orthologues are shared at each node of the taxonomic tree; the shared modifications are shown instead.

In conclusion, the integration of diverse proteomic data sources in the development of PTM Browser has highlighted the need for a consensus on data formats and nomenclature in the field of glycomics. However, despite complex data integration issues, and the fact that computational glycomics is still a relatively 'new' field, the PTM Browser is a powerful tool. It has applications in many areas of molecular biology, including the potential use in an analysis pipeline for the selection of animal models for new drug therapies.

References

- Abbott, A. 2001. And now for the proteome. *Nature* 409, no. 6822: 747. doi:10.1038/35057460.
- Adams, P D. 2001. Regulation of the retinoblastoma tumor suppressor protein by cyclin/cdks. *Biochimica Et Biophysica Acta* 1471, no. 3 (March 21): M123-133.
- Almeida, A., M. Layton, and A. Karadimitris. 2009. Inherited glycosylphosphatidyl inositol deficiency: a treatable CDG. *Biochim Biophys Acta* 1792, no. 9: 874-80. doi:S0925-4439(09)00002-7 [pii] 10.1016/j.bbadis.2008.12.010.
- Amanchy, Ramars, Balamurugan Periaswamy, Suresh Mathivanan, Raghunath Reddy, Sudhir Gopal Tattikota, and Akhilesh Pandey. 2007. A curated compendium of phosphorylation motifs. *Nature Biotechnology* 25, no. 3 (March): 285-286. doi:10.1038/nbt0307-285.
- An, Hyun Joo, John W Froehlich, and Carlito B Lebrilla. 2009. Determination of glycosylation sites and site-specific heterogeneity in glycoproteins. *Current Opinion in Chemical Biology* 13, no. 4 (October): 421-426. doi:10.1016/j.cbpa.2009.07.022.
- Apache-Community. 2010. *The HTTP Apache Server*. <http://httpd.apache.org/>.
- Arrowsmith, C H. 1999. Structure and function in the p53 family. *Cell Death and Differentiation* 6, no. 12 (December): 1169-1173. doi:10.1038/sj.cdd.4400619.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, no. 1: 25--29.
- Attarwala, H. 2010. TGN1412: From Discovery to Disaster. *Journal of Young Pharmacists: JYP* 2, no. 3 (July): 332-336. doi:10.4103/0975-1483.66810.
- Attwood, P V, M.J. Piggott, X L Zu, and P G Besant. 2007. Focus on phosphohistidine. *Amino Acids* 32, no. 1: 145-156. doi:10.1007/s00726-006-0443-6.
- Attwood, P.V., P.G. Besant, and M.J. Piggott. 2010. Focus on phosphoaspartate and phosphoglutamate. *Amino Acids* (September 22). doi:10.1007/s00726-010-0738-5. <http://www.ncbi.nlm.nih.gov/pubmed/20859643>.
- Bairoch, Amos. 2009. UniProt Knowledgebase. Swiss-Prot Protein Knowledgebase. TrEMBL Protein Database. User Manual
- Barrell, D., E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler. 2009. The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 37, no. Database issue: D396-403. doi:gkn803 [pii] 10.1093/nar/gkn803.
- Baumli, Sonja, Jane A Endicott, and Louise N Johnson. 2010. Halogen bonds form the basis for selective P-TEFb inhibition by DRB. *Chemistry & Biology* 17, no. 9 (September 24): 931-936. doi:10.1016/j.chembiol.2010.07.012.
- Belyi, Vladimir A, Prashanth Ak, Elke Markert, Haijian Wang, Wenwei Hu, Anna Puzio-Kuter, and Arnold J Levine. 2010. The origins and evolution of the p53 family of genes. *Cold Spring Harbor Perspectives in Biology* 2, no. 6 (June 1): a001198. doi:10.1101/cshperspect.a001198.
- Benson, Dennis A, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2009. GenBank. *Nucleic Acids Res* 37, no. Database issue: D26--D31.

- Besant, P.G., P.V. Attwood, and M.J. Piggott. 2009. Focus on phosphoarginine and phospholysine. *Current Protein & Peptide Science* 10, no. 6 (December): 536-550.
- Bhagal, N, and R Combes. 2007. Immunostimulatory antibodies: challenging the drug testing paradigm. *Toxicology in Vitro: An International Journal Published in Association with BIBRA* 21, no. 7 (October): 1227-1232. doi:10.1016/j.tiv.2007.02.010.
- Blom, Nikolaj, Thomas Sicheritz-Pont?, Ramneek Gupta, Steen Gammeltoft, and Sren Brunak. 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4, no. 6: 1633--1649.
- Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31, no. 1: 365-70.
- Bohne-Lang, A., E. Lang, T. Forster, and C. W. von der Lieth. 2001. LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* 336, no. 1: 1--11.
- Bordier, Bruno B, Patricia L Marion, Kazuo Ohashi, Mark A Kay, Harry B Greenberg, John L Casey, and Jeffrey S Glenn. 2002. A prenylation inhibitor prevents production of infectious hepatitis delta virus particles. *Journal of Virology* 76, no. 20 (October): 10465-10472.
- Brewer, C F, and L Bhattacharyya. 1986. Specificity of concanavalin A binding to asparagine-linked glycopeptides. A nuclear magnetic relaxation dispersion study. *The Journal of Biological Chemistry* 261, no. 16 (June 5): 7306-7310.
- Brooks, C. L., and W. Gu. 2003. Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr Opin Cell Biol* 15, no. 2: 164-71. doi:S0955067403000036 [pii].
- Bryant, J C, R S Westphal, and B E Wadzinski. 1999. Methylated C-terminal leucine residue of PP2A catalytic subunit is important for binding of regulatory Balph subunit. *The Biochemical Journal* 339 (Pt 2) (April 15): 241-246.
- Cairo-Community. 2010. *Cairo graphics library*. <http://cairographics.org/>.
- Calero, Monica, Catherine Z Chen, Wenyan Zhu, Nena Winand, Karyn A Havas, Penny M Gilbert, Christopher G Burd, and Ruth N Collins. 2003. Dual prenylation is required for Rab protein localization and function. *Molecular Biology of the Cell* 14, no. 5 (May): 1852-1867. doi:10.1091/mbc.E02-11-0707.
- Caragea, C., J. Sinapov, A. Silvescu, D. Dobbs, and V. Honavar. 2007. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics* 8: 438. doi:1471-2105-8-438 [pii] 10.1186/1471-2105-8-438.
- Ceroni, A., K. Maass, H. Geyer, R. Geyer, A. Dell, and S. M. Haslam. 2008. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* 7, no. 4: 1650-9. doi:10.1021/pr7008252.
- Chehab, N H, A Malikzay, E S Stavridi, and T D Halazonetis. 1999. Phosphorylation of Ser-20 mediates stabilization of human p53 in response to DNA damage. *Proceedings of the National Academy of Sciences of the United States of America* 96, no. 24 (November 23): 13777-13782.

- Chellappan, S P, S Hiebert, M Mudryj, J M Horowitz, and J R Nevins. 1991. The E2F transcription factor is a cellular target for the RB protein. *Cell* 65, no. 6 (June 14): 1053-1061.
- Choi, H. S., J. R. Kim, S. W. Lee, and K. H. Cho. 2008. Why have serine/threonine/tyrosine kinases been evolutionarily selected in eukaryotic signaling cascades? *Comput Biol Chem* 32, no. 3: 218-21. doi:S1476-9271(08)00018-2 [pii] 10.1016/j.compbiolchem.2008.02.005.
- Chuikov, S., J. K. Kurash, J. R. Wilson, B. Xiao, N. Justin, G. S. Ivanov, K. McKinney, et al. 2004. Regulation of p53 activity through lysine methylation. *Nature* 432, no. 7015: 353-60. doi:nature03117 [pii] 10.1038/nature03117.
- Clamp, M., B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, and E. S. Lander. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 104, no. 49: 19428-33. doi:0709013104 [pii] 10.1073/pnas.0709013104.
- Clarke, S. 1993. Protein methylation. *Current Opinion in Cell Biology* 5, no. 6 (December): 977-983.
- Clarke, S, J P Vogel, R J Deschenes, and J Stock. 1988. Posttranslational modification of the Ha-ras oncogene protein: evidence for a third class of protein carboxyl methyltransferases. *Proceedings of the National Academy of Sciences of the United States of America* 85, no. 13 (July): 4643-4647.
- Claverol, Stéphane, Odile Burlet-Schiltz, Jean Edouard Gairin, and Bernard Monsarrat. 2003. Characterization of Protein Variants and Post-Translational Modifications: ESI-MSn Analyses of Intact Proteins Eluted from Polyacrylamide Gels. *Molecular & Cellular Proteomics* 2, no. 8: 483-493. doi:10.1074/mcp.T300003-MCP200.
- Coffee, B, F Zhang, S T Warren, and D Reines. 1999. Acetylated histones are associated with FMR1 in normal but not fragile X-syndrome cells. *Nature Genetics* 22, no. 1 (May): 98-101. doi:10.1038/8807.
- Cooper, C. A., M. J. Harrison, M. R. Wilkins, and N. H. Packer. 2001. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res* 29, no. 1: 332-5.
- Cooper, C. A., H. J. Joshi, M. J. Harrison, M. R. Wilkins, and N. H. Packer. 2003. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res* 31, no. 1: 511-3.
- Daubin, V., N. A. Moran, and H. Ochman. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* 301, no. 5634: 829-32. doi:10.1126/science.1086568 301/5634/829 [pii].
- Davidson, Tanja, Erin Beck, Anuradha Ganapathy, Robert Montgomery, Nikhat Zafar, Qi Yang, Ramana Madupu, et al. 2010. The comprehensive microbial resource. *Nucl. Acids Res.* 38, no. suppl_1: D340-345. doi:10.1093/nar/gkp912.
- Davis, M. A., D. Hinerfeld, S. Joseph, Y. H. Hui, N. H. Huang, J. Leszyk, J. Rutherford-Bethard, and S. W. Tam. 2006. Proteomic analysis of rat liver phosphoproteins after treatment with protein kinase inhibitor H89 (N-(2-[p-bromocinnamylamino]-ethyl)-5-isoquinolinesulfonamide). *J Pharmacol Exp Ther* 318, no. 2: 589-95. doi:jpet.105.100032 [pii] 10.1124/jpet.105.100032.

- Delorbe, John E, John H Clements, Benjamin B Whiddon, and Stephen F Martin. 2010. Thermodynamic and Structural Effects of Macrocyclization as a Constraining Method in Protein-Ligand Interactions. *ACS Medicinal Chemistry Letters* 1, no. 8 (November 11): 448-452. doi:10.1021/ml100142y.
- Dessimoz, C., B. Boeckmann, A. C. Roth, and G. H. Gonnet. 2006. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res* 34, no. 11: 3309-16. doi:34/11/3309 [pii] 10.1093/nar/gkl433.
- Diella, Francesca, Scott Cameron, Christine Gemnd, Rune Linding, Allegra Via, Bernhard Kuster, Thomas Sicheritz-Pont?, Nikolaj Blom, and Toby J Gibson. 2004. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 5: 79.
- Diella, Francesca, Cathryn M Gould, Claudia Chica, Allegra Via, and Toby J Gibson. 2008a. Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res* 36, no. Database issue: D240--D244.
- Diella, Francesca, Niall Haslam, Claudia Chica, Aidan Budd, Sushama Michael, Nigel P Brown, Gilles Trave, and Toby J Gibson. 2008b. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci* 13: 6580--6603.
- Doubet, S., K. Bock, D. Smith, A. Darvill, and P. Albersheim. 1989. The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 14, no. 12: 475--477.
- Dowell, R. D., R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein. 2001. The distributed annotation system. *BMC Bioinformatics* 2: 7.
- Downward, Julian. 2003. Targeting RAS signalling pathways in cancer therapy. *Nature Reviews. Cancer* 3, no. 1: 11-22. doi:10.1038/nrc969.
- Dumaz, N, D M Milne, and D W Meek. 1999. Protein kinase CK1 is a p53-threonine 18 kinase which requires prior phosphorylation of serine 15. *FEBS Letters* 463, no. 3 (December 17): 312-316.
- Durek, P., R. Schmidt, J. L. Heazlewood, A. Jones, D. MacLean, A. Nagel, B. Kersten, and W. X. Schulze. 2010. PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res* 38, no. Database issue: D828-34. doi:gkp810 [pii] 10.1093/nar/gkp810.
- Dziarski, R. 2003. Recognition of bacterial peptidoglycan by the innate immune system. *Cell Mol Life Sci* 60, no. 9: 1793-804. doi:10.1007/s00018-003-3019-6.
- Eberharther, Anton, and Peter B Becker. 2002. Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics. *EMBO Reports* 3, no. 3 (March): 224-229. doi:10.1093/embo-reports/kvf053.
- Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* 14, no. 9: 755-63. doi:btb114 [pii].
- Elsik, Christine G., Ross L. Tellam, and Kim C. Worley, with Sequencing, The Bovine Genome, Analysis Consortium. 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* 324, no. 5926: 522-528. doi:10.1126/science.1169588.
- Fan, Xiaolian, Yi-Min She, Rick D Bagshaw, John W Callahan, Harry Schachter, and Don J Mahuran. 2004. A method for proteomic identification of membrane-bound proteins containing Asn-linked oligosaccharides. *Analytical*

- Biochemistry* 332, no. 1 (September 1): 178-186.
doi:10.1016/j.ab.2004.05.038.
- Farriol-Mathis, Nathalie, John S Garavelli, Brigitte Boeckmann, S?erine Duvaud, Elisabeth Gasteiger, Alain Gateau, Anne-Lise Veuthey, and Amos Bairoch. 2004. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* 4, no. 6: 1537--1550.
- Fellinger, Michael. 2010. *Ramaze - The Modular Web Framework*.
<http://ramaze.net/>.
- Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166.
- Feng, Zukang, Li Chen, Himabindu Maddula, Ozgur Akcan, Rose Oughtred, Helen M Berman, and John Westbrook. 2004. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics (Oxford, England)* 20, no. 13 (September 1): 2153-2155. doi:10.1093/bioinformatics/bth214.
- Ferguson, M. A., and A. F. Williams. 1988. Cell-surface anchoring of proteins via glycosyl-phosphatidylinositol structures. *Annu Rev Biochem* 57: 285-320.
doi:10.1146/annurev.bi.57.070188.001441.
- Filhol, O, J Baudier, C Delphin, P Loue-Mackebach, E M Chambaz, and C Cochet. 1992. Casein kinase II and the tumor suppressor protein P53 associate in a molecular complex that is negatively regulated upon P53 phosphorylation. *The Journal of Biological Chemistry* 267, no. 29 (October 15): 20577-20583.
- Finn, R. D., J. Mistry, J. Tate, P. Coghill, A. Heger, J. E. Pollington, O. L. Gavin, et al. 2010. The Pfam protein families database. *Nucleic Acids Res* 38, no. Database issue: D211-22. doi:gkp985 [pii] 10.1093/nar/gkp985.
- Finn, R. D., J. Tate, J. Mistry, P. C. Coghill, S. J. Sammut, H. R. Hotz, G. Ceric, et al. 2008. The Pfam protein families database. *Nucleic Acids Res* 36, no. Database issue: D281-8. doi:gkm960 [pii] 10.1093/nar/gkm960.
- Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* 19, no. 2: 99-113.
- Fitch, W.M. 2000. Homology a personal view on some of the problems. *Trends Genet* 16, no. 5: 227-31. doi:S0168-9525(00)02005-9 [pii].
- Fleischmann, Wolfgang, Alexandre Gattiker, Henning Hermjakob, Eric Jain, and Paul Kersey. 2007. *Swissknife. An object-oriented Perl library to handle Swiss-Prot entries*. 1., European Bioinformatics Institute, 2., (Swiss Institute of Bioinformatics). <http://swissknife.sourceforge.net/docs/>.
- Flicek, P., B. L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, et al. 2010. Ensembl's 10th year. *Nucleic Acids Res* 38, no. Database issue: D557-62. doi:gkp972 [pii] 10.1093/nar/gkp972.
- Frank, M., and S. Schloissnig. 2010. Bioinformatics and molecular modeling in glycobiology. *Cell Mol Life Sci* 67, no. 16: 2749-72. doi:10.1007/s00018-010-0352-4.
- Fuchs, Stephen M, Krzysztof Krajewski, Richard W Baker, Victoria L Miller, and Brian D Strahl. 2011. Influence of combinatorial histone modifications on antibody and effector protein recognition. *Current Biology: CB* 21, no. 1: 53-58. doi:10.1016/j.cub.2010.11.058.
- Garavelli, John S. 2003. The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res* 31, no. 1: 499--501.
- Gnad, F., S. Ren, J. Cox, J. V. Olsen, B. Macek, M. Orosi, and M. Mann. 2007. PHOSIDA (phosphorylation site database): management, structural and

- evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8, no. 11: R250. doi:gb-2007-8-11-r250 [pii] 10.1186/gb-2007-8-11-r250.
- Gnad, Florian, Jeremy Gunawardena, and Matthias Mann. 2011. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Research* 39, no. Database issue: D253-260. doi:10.1093/nar/gkq1159.
- Goldberg, David, Mark Sutton-Smith, James Paulson, and Anne Dell. 2005. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* 5, no. 4 (March): 865-875. doi:10.1002/pmic.200401071.
- Grillo, M A, and S Colombatto. 2005. S-adenosylmethionine and protein methylation. *Amino Acids* 28, no. 4 (June): 357-362. doi:10.1007/s00726-005-0197-6.
- Gruber, Thomas R. 1993. *A Translation Approach to Portable Ontology Specifications*. Stanford University.
- Gupta, N., S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, et al. 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* 17, no. 9: 1362-77. doi:gr.6427907 [pii] 10.1101/gr.6427907.
- Gupta, R., H. Birch, K. Rapacki, S. Brunak, and J. E. Hansen. 1999. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 27, no. 1: 370--372.
- Hakomori, S. 2002. Glycosylation defining cancer malignancy: new wine in an old bottle. *Proc Natl Acad Sci U S A* 99, no. 16: 10231-3. doi:10.1073/pnas.172380699 172380699 [pii].
- Hansen, J. E., O. Lund, J. O. Nielsen, and S. Brunak. 1996. O-GLYCBASE: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 24, no. 1: 248--252.
- Harb, George, Rupangi C Vasavada, David Cobrinik, and Andrew F Stewart. 2009. The retinoblastoma protein and its homolog p130 regulate the G1/S transition in pancreatic beta-cells. *Diabetes* 58, no. 8 (August): 1852-1862. doi:10.2337/db08-0759.
- Hashimoto, K., S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki, and M. Kanehisa. 2006. KEGG as a glycome informatics resource. *Glycobiology* 16, no. 5: 63R-70R. doi:cwj010 [pii] 10.1093/glycob/cwj010.
- Hayes, Catherine A, Niclas G Karlsson, Weston B Struwe, Frederique Lisacek, Pauline M Rudd, Nicole H Packer, and Matthew P Campbell. 2011. UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics (Oxford, England)* 27, no. 9 (May 1): 1343-1344. doi:10.1093/bioinformatics/btr137.
- He, X., and J. Zhang. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, no. 2: 1157-64. doi:genetics.104.037051 [pii] 10.1534/genetics.104.037051.
- Heazlewood, J. L., P. Durek, J. Hummel, J. Selbig, W. Weckwerth, D. Walther, and W. X. Schulze. 2008. PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res* 36, no. Database issue: D1015-21. doi:gkm812 [pii] 10.1093/nar/gkm812.
- Hellen, Beth, Ruth Spriggs, David Damerell, and Sue Jones. 2008. A combined Transcription factor prediction tool.

- Herget, S, R Ranzinger, K Maass, and C-W V D Lieth. 2008. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydrate Research* 343, no. 12 (August 11): 2162-2171. doi:10.1016/j.carres.2008.03.011.
- Higashimoto, Y, S Saito, X H Tong, A Hong, K Sakaguchi, E Appella, and C W Anderson. 2000. Human p53 is phosphorylated on serines 6 and 9 in response to DNA damage-inducing agents. *The Journal of Biological Chemistry* 275, no. 30 (July 28): 23199-23203. doi:10.1074/jbc.M002674200.
- Hirao, A, Y Y Kong, S Matsuoka, A Wakeham, J Ruland, H Yoshida, D Liu, S J Elledge, and T W Mak. 2000. DNA damage-induced activation of p53 by the checkpoint kinase Chk2. *Science (New York, N.Y.)* 287, no. 5459 (March 10): 1824-1827.
- Hitchen, P. G., and A. Dell. 2006. Bacterial glycoproteomics. *Microbiology* 152, no. Pt 6: 1575-80. doi:152/6/1575 [pii] 10.1099/mic.0.28859-0.
- Honda, R., H. Tanaka, and H. Yasuda. 1997. Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. *FEBS Lett* 420, no. 1: 25-7. doi:S0014-5793(97)01480-4 [pii].
- Hornbeck, Peter V, Indy Chabra, Jon M Kornhauser, Elzbieta Skrzypek, and Bin Zhang. 2004. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 4, no. 6 (June): 1551-1561. doi:10.1002/pmic.200300772.
- Huang, Hsien-Da, Tzong-Yi Lee, Shih-Wei Tzeng, and Jorng-Tzong Horng. 2005a. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res* 33, no. Web Server issue: W226--W229.
- Huang, Hsien-Da, Tzong-Yi Lee, Shih-Wei Tzeng, Li-Cheng Wu, Jorng-Tzong Horng, Ann-Ping Tsou, and Kuan-Tsae Huang. 2005b. Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J Comput Chem* 26, no. 10: 1032--1041.
- Hurles, M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol* 2, no. 7: E206. doi:10.1371/journal.pbio.0020206.
- Hwang, Cheol-Sang, Anna Shemorry, and Alexander Varshavsky. 2010. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science (New York, N.Y.)* 327, no. 5968 (February 19): 973-977. doi:10.1126/science.1183147.
- Initiative, The Arabidopsis Genome. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, no. 6814: 796-815.
- JQuery-Community. 2010. *jQuery - write less do more*. <http://jquery.com/>.
- IRuby-Community. 2010. *IRuby*. <http://jruby.org>.
- Jaeken, J., and G. Matthijs. 2007. Congenital disorders of glycosylation: a rapidly expanding disease family. *Annu Rev Genomics Hum Genet* 8: 261-78. doi:10.1146/annurev.genom.8.080706.092327.
- Julenius, K., A. Molgaard, R. Gupta, and S. Brunak. 2005. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* 15, no. 2: 153-64. doi:10.1093/glycob/cwh151 cwh151 [pii].
- KEGG. 2010. *KEGG API*. <http://www.genome.jp/kegg/soap/>.
- Kühlberg, Axel, Mark Haid, and Sabine Metzger. 2010. Characterization of O-phosphohydroxyproline in rat {alpha}-crystallin A. *The Journal of Biological*

- Chemistry* 285, no. 41 (October 8): 31484-31490.
doi:10.1074/jbc.M109.035428.
- Khorchid, A., and M. Ikura. 2006. Bacterial histidine kinase as signal sensor and transducer. *Int J Biochem Cell Biol* 38, no. 3: 307-12. doi:S1357-2725(05)00260-8 [pii] 10.1016/j.biocel.2005.08.018.
- Kim, Sunshin, Kwang Jung, and Keun Ryu. 2006. Automatic Orthologous-Protein-Clustering from Multiple Complete-Genomes by the Best Reciprocal BLAST Hits. In *Data Mining for Biomedical Applications*, 60-70.
http://dx.doi.org/10.1007/11691730_7.
- Klotz, A V, and A N Glazer. 1987. gamma-N-methylasparagine in phycobiliproteins. Occurrence, location, and biosynthesis. *The Journal of Biological Chemistry* 262, no. 36 (December 25): 17350-17355.
- Koonin, Eugene V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39: 309--338.
- Kuzniar, Arnold, Roeland C H J van Ham, S  ndor Pongor, and Jack A M Leunissen. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* 24, no. 11: 539--551.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, no. 6822: 860--921.
- Lapko, Veniamin N, David L Smith, and Jean B Smith. 2002. S-methylated cysteines in human lens gamma S-crystallins. *Biochemistry* 41, no. 50 (December 17): 14645-14651.
- Lee, Tzong-Yi, Hsien-Da Huang, Jui-Hung Hung, Hsi-Yuan Huang, Yuh-Shyong Yang, and Tzu-Hao Wang. 2006. dbPTM: an information repository of protein post-translational modification. *Nucl. Acids Res.* 34, no. suppl\1: D622--627.
- Linder, Maurine E, and Robert J Deschenes. 2003. New insights into the mechanisms of protein palmitoylation. *Biochemistry* 42, no. 15: 4311--4320.
- Liolios, K., I. M. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, and N. C. Kyrpides. 2010. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 38, no. Database issue: D346-54. doi:gkp848 [pii] 10.1093/nar/gkp848.
- Liu, Qiyuan, Satoshi Kaneko, Lin Yang, Richard I Feldman, Santo V Nicosia, Jiandong Chen, and Jin Q Cheng. 2004. Aurora-A abrogation of p53 DNA binding and transactivation activity by phosphorylation of serine 215. *The Journal of Biological Chemistry* 279, no. 50 (December 10): 52175-52182. doi:10.1074/jbc.M406802200.
- Lord, P. W., R. D. Stevens, A. Brass, and C. A. Goble. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, no. 10: 1275-83.
- Loss, Alexander, Peter Bunsmann, Andreas Bohne, Annika Loss, Eberhard Schwarzer, Elke Lang, and Claus-W. von der Lieth. 2002. SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 30, no. 1: 405--408.
- Lutteke, T., M. Frank, and C. W. von der Lieth. 2004. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures.

- Carbohydr Res* 339, no. 5: 1015-20. doi:10.1016/j.carres.2003.09.038 S0008621503005755 [pii].
- Ma, Bruce Y, Sebastian A Mikolajczak, Tetsuya Yoshida, Ryoko Yoshida, David J Kelvin, and Atsuo Ochi. 2004. CD28 T cell costimulatory receptor function is negatively regulated by N-linked carbohydrates. *Biochemical and Biophysical Research Communications* 317, no. 1 (April 23): 60-67. doi:10.1016/j.bbrc.2004.03.012.
- Macek, B., F. Gnad, B. Soufi, C. Kumar, J. V. Olsen, I. Mijakovic, and M. Mann. 2008. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics* 7, no. 2: 299-307. doi:M700311-MCP200 [pii] 10.1074/mcp.M700311-MCP200.
- Marechal, V, B Elenbaas, L Taneyhill, J Piette, M Mechali, J C Nicolas, A J Levine, and J Moreau. 1997. Conservation of structural domains and biochemical activities of the MDM2 protein from *Xenopus laevis*. *Oncogene* 14, no. 12 (March 27): 1427-1433. doi:10.1038/sj.onc.1200967.
- Martin, Andrew C R. 2005. Mapping PDB chains to UniProtKB entries. *Bioinformatics* 21, no. 23: 4297--4301.
- Matsuoka, Shuhei, Bryan A Ballif, Agata Smogorzewska, E Robert McDonald 3rd, Kristen E Hurov, Ji Luo, Corey E Bakalarski, et al. 2007. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science (New York, N.Y.)* 316, no. 5828 (May 25): 1160-1166. doi:10.1126/science.1140321.
- Maurer-Stroh, Sebastian, and Frank Eisenhaber. 2005. Refinement and prediction of protein prenylation motifs. *Genome Biol* 6, no. 6: R55.
- McAdams, Kenneth, Eric S Casper, R Matthew Haas, Bernard D Santarsiero, Aimee L Eggler, Andrew Mesecar, and Christopher J Halkides. 2008. The structures of T87I phosphono-CheY and T87I/Y106W phosphono-CheY help to explain their binding affinities to the FliM and CheZ peptides. *Archives of Biochemistry and Biophysics* 479, no. 2 (November 15): 105-113. doi:10.1016/j.abb.2008.08.019.
- Meckler, Xavier, Jelita Roseman, Pritam Das, Haipeng Cheng, Susan Pei, Marcia Keat, Breanne Kassarian, Todd E Golde, Angèle T Parent, and Gopal Thinakaran. 2010. Reduced Alzheimer's disease β -amyloid deposition in transgenic mice expressing S-palmitoylation-deficient APH1aL and nicastrin. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30, no. 48 (December 1): 16160-16169. doi:10.1523/JNEUROSCI.4436-10.2010.
- Merkeev, I. V., P. S. Novichkov, and A. A. Mironov. 2006. PHOG: a database of supergenomes built from proteome complements. *BMC Evol Biol* 6: 52. doi:1471-2148-6-52 [pii] 10.1186/1471-2148-6-52.
- Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11, no. 1: 31-46. doi:nrg2626 [pii] 10.1038/nrg2626.
- Miller, Ingrid, Johanne Crawford, and Elisabetta Gianazza. 2006. Protein stains for proteomic applications: which, when, why? *Proteomics* 6, no. 20 (October): 5385-5408. doi:10.1002/pmic.200600323.
- Minamoto, T, T Buschmann, H Habelhah, E Matusevich, H Tahara, A L Boerresen-Dale, C Harris, D Sidransky, and Z Ronai. 2001. Distinct pattern of p53 phosphorylation in human tumors. *Oncogene* 20, no. 26 (June 7): 3341-3347. doi:10.1038/sj.onc.1204458.

- Montecchi-Palazzi, Luisa, Ron Beavis, Pierre-Alain Binz, Robert J Chalkley, John Cottrell, David Creasy, Jim Shofstahl, Sean L Seymour, and John S Garavelli. 2008. The PSI-MOD community standard for representation of protein modification data. *Nat Biotechnol* 26, no. 8: 864--866.
- NCBI-Entrez. 2010. *NCBI Entrez Utilities Web Service*. http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html.
- Nagashima, Kumiko, Stuart D Shumway, Sriram Sathyanarayanan, Albert H Chen, Brian Dolinski, Youyuan Xu, Heike Keilhack, et al. 2011. Genetic and Pharmacological Inhibition of PDK1 in Cancer Cells: CHARACTERIZATION OF A SELECTIVE ALLOSTERIC KINASE INHIBITOR. *The Journal of Biological Chemistry* 286, no. 8 (February 25): 6433-6448. doi:10.1074/jbc.M110.156463.
- Nakahara, Taku, Ryo Hashimoto, Hiroaki Nakagawa, Kenji Monde, Nobuaki Miura, and Shin-Ichiro Nishimura. 2008. Glycoconjugate Data Bank: Structures--an annotated glycan structure database and N-glycan primary structure verification service. *Nucleic Acids Res* 36, no. Database issue: D368--D371.
- O'Brien, K. P., I. Westerlund, and E. L. Sonnhammer. 2004. OrthoDisease: a database of human disease orthologs. *Hum Mutat* 24, no. 2: 112-9. doi:10.1002/humu.20068.
- O'Donovan, C., M. J. Martin, E. Glemet, J. J. Codani, and R. Apweiler. 1999. Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics* 15, no. 3: 258-9. doi:btc050 [pii].
- Oracle. 2010. *The MySQL RDMS*. Oracle. <http://www.mysql.com/>.
- Pankow, Sandra, and Casimir Bamberger. 2007. The p53 tumor suppressor-like protein nvp63 mediates selective germ cell death in the sea anemone *Nematostella vectensis*. *PloS One* 2, no. 9: e782. doi:10.1371/journal.pone.0000782.
- Parham, P. 1996. Functions for MHC class I carbohydrates inside and outside the cell. *Trends Biochem Sci* 21, no. 11: 427-33. doi:S0968-0004(96)10053-0 [pii].
- Patwardhan, Parag, and Marilyn D Resh. 2010. Myristoylation and membrane binding regulate c-Src stability and kinase activity. *Molecular and Cellular Biology* 30, no. 17 (September): 4094-4107. doi:10.1128/MCB.00246-10.
- Peek, K, R M Daniel, C Monk, L Parker, and T Coolbear. 1992. Purification and characterization of a thermostable proteinase isolated from *Thermus* sp. strain Rt41A. *European Journal of Biochemistry / FEBS* 207, no. 3 (August 1): 1035-1044.
- Perez, S., and B. Mulloy. 2005. Prospects for glycoinformatics. *Curr Opin Struct Biol* 15, no. 5: 517-24. doi:S0959-440X(05)00153-3 [pii] 10.1016/j.sbi.2005.08.005.
- Pfam-Consortium. 2010. Pfam online documentation. <http://pfam.sanger.ac.uk>.
- Philips, Mark R. 2004. Methotrexate and Ras methylation: a new trick for an old drug? *Science's STKE: Signal Transduction Knowledge Environment* 2004, no. 225 (March 23): pe13. doi:10.1126/stke.2252004pe13.
- Polevoda, B, J Norbeck, H Takakura, A Blomberg, and F Sherman. 1999. Identification and specificities of N-terminal acetyltransferases from *Saccharomyces cerevisiae*. *The EMBO Journal* 18, no. 21 (November 1): 6155-6168. doi:10.1093/emboj/18.21.6155.

- Prince, John T, Mark W Carlson, Rong Wang, Peng Lu, and Edward M Marcotte. 2004. The need for a public proteomics repository. *Nat Biotechnol* 22, no. 4: 471--472.
- Pruess, Manuela, Paul Kersey, and Rolf Apweiler. 2005. The Integr8 project--a resource for genomic and proteomic data. *In Silico Biol* 5, no. 2: 179--185.
- Punternvoll, Pal, Rune Linding, Christine Gemund, Sophie Chabanis-Davidson, Morten Mattingsdal, Scott Cameron, David M A Martin, et al. 2003. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 31, no. 13: 3625--3630.
- Putnam, Nicholas H., Mansi Srivastava, Uffe Hellsten, Bill Dirks, Jarrod Chapman, Asaf Salamov, Astrid Terry, et al. 2007. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* 317, no. 5834: 86-94. doi:10.1126/science.1139158.
- Raju, R V, T N Moyana, and R K Sharma. 1997. N-Myristoyltransferase overexpression in human colorectal adenocarcinomas. *Experimental Cell Research* 235, no. 1 (August 25): 145-154. doi:10.1006/excr.1997.3679.
- Raman, R., M. Venkataraman, S. Ramakrishnan, W. Lang, S. Raguram, and R. Sasisekharan. 2006. Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology* 16, no. 5: 82R-90R. doi:cwj080 [pii] 10.1093/glycob/cwj080.
- Remm, M., C. E. Storm, and E. L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314, no. 5: 1041-52. doi:10.1006/jmbi.2000.5197 S0022-2836(00)95197-0 [pii].
- Rholam, M., P. Nicolas, and P. Cohen. 1986. Precursors for peptide hormones share common secondary structures forming features at the proteolytic processing sites. *FEBS Lett* 207, no. 1: 1-6. doi:0014-5793(86)80002-3 [pii].
- Robinson, L. J., and T. Michel. 1995. Mutagenesis of palmitoylation sites in endothelial nitric oxide synthase identifies a novel motif for dual acylation and subcellular targeting. *Proc Natl Acad Sci U S A* 92, no. 25: 11776-80.
- Rodriguez, M. S., J. M. Desterro, S. Lain, C. A. Midgley, D. P. Lane, and R. T. Hay. 1999. SUMO-1 modification activates the transcriptional response of p53. *EMBO J* 18, no. 22: 6455-61. doi:10.1093/emboj/18.22.6455.
- Roskoski, Robert. 2003. Protein prenylation: a pivotal posttranslational process. *Biochem Biophys Res Commun* 303, no. 1: 1-7.
- Ryle, A. P., F. Sanger, L. F. Smith, and R. Kitai. 1955. The disulphide bonds of insulin. *Biochem J* 60, no. 4: 541-56.
- Sahoo, Satya S, Christopher Thomas, Amit Sheth, Cory Henson, and William S York. 2005. GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydrate Research* 340, no. 18 (December 30): 2802-2807. doi:10.1016/j.carres.2005.09.019.
- Sanger, F., and E.O. Thompson. 1953a. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J* 53, no. 3: 353-66.
- . 1953b. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem J* 53, no. 3: 366-74.

- Sanger, F., and H. Tuppy. 1951a. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J* 49, no. 4: 481-90.
- . 1951b. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J* 49, no. 4: 463-81.
- Sayers, Eric W, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37, no. Database issue: D5--15.
- Selmer, T, J Kahnt, M Goubeaud, S Shima, W Grabarse, U Ermler, and R K Thauer. 2000. The biosynthesis of methylated amino acids in the active site region of methyl-coenzyme M reductase. *The Journal of Biological Chemistry* 275, no. 6 (February 11): 3755-3760.
- Senger, Ryan S, and M. Nazmul Karim. 2008. Prediction of N-linked glycan branching patterns using artificial neural networks. *Math Biosci* 211, no. 1: 89--104.
- Sevier, C. S., and C. A. Kaiser. 2002. Formation and transfer of disulphide bonds in living cells. *Nat Rev Mol Cell Biol* 3, no. 11: 836-47. doi:10.1038/nrm954 nrm954 [pii].
- Shieh, S. Y., J. Ahn, K. Tamai, Y. Taya, and C. Prives. 2000. The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites. *Genes Dev* 14, no. 3: 289-300.
- Sigrist, C. J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3, no. 3: 265-74.
- Sigrist, C. J., E. De Castro, P. S. Langendijk-Genevaux, V. Le Saux, A. Bairoch, and N. Hulo. 2005. ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* 21, no. 21: 4060-6. doi:bt614 [pii] 10.1093/bioinformatics/bti614.
- Sigrist, C.J., Lorenzo Cerutti, Edouard de Castro, Petra S Langendijk-Genevaux, Virginie Bulliard, Amos Bairoch, and Nicolas Hulo. 2010. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research* 38, no. Database issue: D161-166. doi:10.1093/nar/gkp885.
- Smith, Barry, Werner Ceusters, Bert Klagges, Jacob Kohler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. 2005. Relations in biomedical ontologies. *Genome Biol* 6, no. 5: R46.
- Smith, J. M. 1970. Natural selection and the concept of a protein space. *Nature* 225, no. 5232: 563-4.
- Soppa, Jörg. 2010. Protein acetylation in archaea, bacteria, and eukaryotes. *Archaea (Vancouver, B.C.)* 2010. doi:10.1155/2010/820681. <http://www.ncbi.nlm.nih.gov/pubmed/20885971>.
- Spiro, R. G. 2002. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 12, no. 4: 43R-56R.
- Steinberg, Thomas H. 2009. Protein gel staining methods: an introduction and overview. *Methods in Enzymology* 463: 541-563. doi:10.1016/S0076-6879(09)63031-7.

- Steiner, D. F., and P. E. Oyer. 1967. The biosynthesis of insulin and a probable precursor of insulin by a human islet cell adenoma. *Proc Natl Acad Sci U S A* 57, no. 2: 473-80.
- Stevens, R, A Rector, and D Hull. 2010. What is an ontology? | Ontogenesis. <http://ontogenesis.knowledgeblog.org/66>.
- Strecker, Thomas, Anna Maisa, Stephane Daffis, Robert Eichler, Oliver Lenz, and Wolfgang Garten. 2006. The role of myristoylation in the membrane association of the Lassa virus matrix protein Z. *Virology Journal* 3: 93. doi:10.1186/1743-422X-3-93.
- Suzek, Baris E., Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, no. 10: 1282-1288. doi:10.1093/bioinformatics/btm098.
- Szymanski, C M, R Yao, C P Ewing, T J Trust, and P Guerry. 1999. Evidence for a system of general protein glycosylation in *Campylobacter jejuni*. *Molecular Microbiology* 32, no. 5 (June): 1022-1030.
- Takasaki, S., K. Yamashita, and A. Kobata. 1978. The sugar chain structures of ABO blood group active glycoproteins obtained from human erythrocyte membrane. *J Biol Chem* 253, no. 17: 6086-91.
- Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41. doi:10.1186/1471-2105-4-41 1471-2105-4-41 [pii].
- Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28, no. 1: 33-6. doi:gkd013 [pii].
- Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. *Science* 278, no. 5338: 631-7.
- Taylor, Paul B., and Elizabeth A. Cook. 1981. Myocardial histone acetylation. *Life Sciences* 28, no. 21 (May 21): 2403-2410. doi:doi: DOI: 10.1016/0024-3205(81)90507-5.
- Thompson, J D, D G Higgins, and T J Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22, no. 22 (November 11): 4673-4680.
- Tyagi, Alpana, Chapla Agarwal, and Rajesh Agarwal. 2002. Inhibition of retinoblastoma protein (Rb) phosphorylation at serine sites and an increase in Rb-E2F complex formation by silibinin in androgen-dependent human prostate carcinoma LNCaP cells: role in prostate cancer prevention. *Molecular Cancer Therapeutics* 1, no. 7 (May): 525-532.
- UniProt Consortium. 2009. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 37, no. Database issue: D169--D174.
- UniProtKB. 2010. UniProtKB/Swiss-Prot Protein Knowledgebase - Swiss-Prot headline. <http://expasy.org/sprot/relnotes/spwrnew.html>.
- Van Damme, Petra, Jozef Van Damme, Hans Demol, An Staes, Joël Vandekerckhove, and Kris Gevaert. 2009. A review of COFRADIC techniques targeting protein N-terminal acetylation. *BMC Proceedings* 3 Suppl 6: S6. doi:10.1186/1753-6561-3-S6-S6.

- Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. The sequence of the human genome. *Science* 291, no. 5507: 1304--1351.
- Vetrivel, Kulandaivelu S, Xavier Meckler, Ying Chen, Phuong D Nguyen, Nabil G Seidah, Robert Vassar, Philip C Wong, Masaki Fukata, Maria Z Kounnas, and Gopal Thinakaran. 2009. Alzheimer disease Abeta production in the absence of S-palmitoylation-dependent targeting of BACE1 to lipid rafts. *The Journal of Biological Chemistry* 284, no. 6 (February 6): 3793-3803. doi:10.1074/jbc.M808920200.
- Walsh, Christopher T. 2006. *Posttranslational Modification of Proteins Expanding Nature's Inventory*. Roberts and Company.
- Wilkinson, M. D., and M. Links. 2002. BioMOBY: an open source biological web services proposal. *Brief Bioinform* 3, no. 4: 331-41.
- Wilson, Keith. 2000. *Principles & techniques of practical biochemistry*. 5th ed. Cambridge: Cambridge University Press.
- Wright, Megan H, William P Heal, David J Mann, and Edward W Tate. 2009. Protein myristoylation in health and disease. *Journal of Chemical Biology* (November 7). doi:10.1007/s12154-009-0032-8. <http://www.ncbi.nlm.nih.gov/pubmed/19898886>.
- Wu, Xiaomei, Lei Zhu, Jie Guo, Da-Yong Zhang, and Kui Lin. 2006. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res* 34, no. 7: 2137--2150.
- Xiang, Tao, Qun Liu, Ashley M Deacon, Matthew Koshy, Irina A Kriksunov, Xin Gen Lei, Quan Hao, and Daniel J Thiel. 2004. Crystal structure of a heat-resilient phytase from *Aspergillus fumigatus*, carrying a phosphorylated histidine. *Journal of Molecular Biology* 339, no. 2 (May 28): 437-445. doi:10.1016/j.jmb.2004.03.057.
- Xirodimas, D. P., M. K. Saville, J. C. Bourdon, R. T. Hay, and D. P. Lane. 2004. Mdm2-mediated NEDD8 conjugation of p53 inhibits its transcriptional activity. *Cell* 118, no. 1: 83-97. doi:10.1016/j.cell.2004.06.016 S0092867404005859 [pii].
- Yamamoto, F., H. Clausen, T. White, J. Marken, and S. Hakomori. 1990. Molecular genetic basis of the histo-blood group ABO system. *Nature* 345, no. 6272: 229-33. doi:10.1038/345229a0.
- Yazer, M. H., and M. L. Olsson. 2008. The O2 allele: questioning the phenotypic definition of an ABO allele. *Immunohematology* 24, no. 4: 138-47.
- Yonemoto, W, M L McGlone, and S S Taylor. 1993. N-myristylation of the catalytic subunit of cAMP-dependent protein kinase conveys structural stability. *The Journal of Biological Chemistry* 268, no. 4 (February 5): 2348-2352.
- York, Will. 2011. GLYDE-II. <http://glycomics.ccr.cu.edu/core4/informatics-glyde-ii.html>.
- Yu, Jun, Songnian Hu, Jun Wang, Gane Ka-Shu Wong, Songgang Li, Bin Liu, Yajun Deng, et al. 2002. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. indica). *Science* 296, no. 5565: 79-92. doi:10.1126/science.1068037.
- Yurist-Doutsch, Sophie, Bonnie Chaban, David J VanDyke, Ken F Jarrell, and Jerry Eichler. 2008. Sweet to the extreme: protein glycosylation in Archaea. *Molecular Microbiology* 68, no. 5 (June): 1079-1084. doi:10.1111/j.1365-2958.2008.06224.x.

Zhang, Liangyi, and James P Reilly. 2009. Extracting both peptide sequence and glycan structural information by 157 nm photodissociation of N-linked glycopeptides. *Journal of Proteome Research* 8, no. 2 (February): 734-742. doi:10.1021/pr800766f.

Appendix

Appendix 1: p53 family UniProtKB list

Q9W678, P67938, P67939, Q29537, Q9WUR6, P13481, O09185, P79734, Q8SPZ3, P41685, P10360, P04637, O93379, P56423, P61260, P56424, O36006, Q00366, A7SFL1, P25035, Q95330, P79820, P51664, O12946, P10361, Q9TUB2, Q9W679, Q9TTA1, P07193, O57538, Q92143, Q9JJP6, O88898, Q9H3D4, Q8JHZ6, Q98SW0, Q9DEC7, Q9JJP2, Q9XSK8, O15350, Q5KQU6, Q4SF81

Appendix 2: Source code

The following is a list of some of the more important programs that have been created as part of this thesis. The PTM Browser web site should be consulted for a list of the latest versions of these programs and other programs. All of the source code that has been written as part of this project is available for download from a number of repositories. Instructions pertaining to the download and setup of this source code can be found at the following URL:

<http://wiki.ptmbrowser.org/index.php/BBP_Setup>

Program: org.drd20.bioinformatics.database.ptmdb.psiModAutomate.sh

Input: \$BBP/org/drd20/bioinformatics/database/ptmdb/PSI-MOD.obo (static)

(latest version: <http://psidev.sourceforge.net/mod/data/PSI-MOD.obo>)

Description

This script is responsible for parsing an OBO formatted PSI-MOD ontology file. Which can be downloaded from the following URL

Program: org.drd20.bioinformatics.database.ptmdb.PtmDbErd (class)

Description

Class provides methods to add new PTM annotations to the PTMDB.

Program: ptmDB/ncbiEntrezTaxonomy/namesToMySQL.pl

Input: /tmp/names.dmp (static)

(latest version: ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz)

Description

This script imports the NCBI Entrez Taxonomy name dump file into the PTMDB.

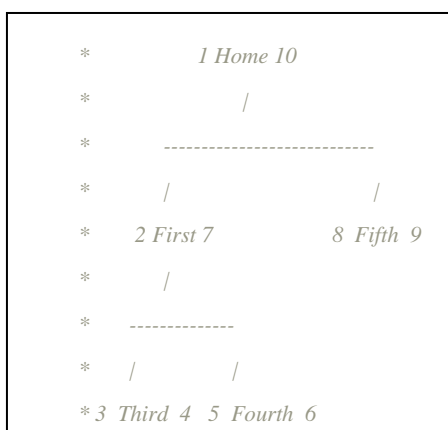
Program: ptmDB/ncbiEntrezTaxonomy/createRelationshipTable.pl

Input: /tmp/nodes.dmp (static)

(latest version: ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz)

Description

This script imports the NCBI Entrez Taxonomy tree from the nodes dump file into the PTMDB. The most obvious way of storing the taxonomic tree in the database is to connect each node to its immediate ancestor. However this makes the process of selecting all the descendants of a particular node very slow as it involves a great deal of graph traversing. The solution that has been used in the PTMDB is to traverse the taxonomic tree and store every possible ancestor/descendent relationship that is observed (i.e. there is one row that connects *H. sapiens* to Mammalia and another that connects *H. sapiens* to Eukaryota). An alternative would have been to create a pre-order tree, like that shown below, that gives each node a left and right index.



By discovering the left and right index of the ancestor node it is then possible to obtain all descendants by asking for nodes with a left index which is > the ancestors index but whose right index is less than that of the ancestor.

Program: ptmDB/Main/UniprotUpload/uniprotToDatabaseErd.pl

Input:

- t Database type [TREMBL or SW]
- d Location of UniProtKB DAT file
- e Erase tables common to Swiss-Prot and TrEMBL [0|1]
- E Include additional tables (creates some additional referencing tables)
- P Import annotations into the PTMDB

Latest Swiss-Prot DAT file:

http://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz

Latest TrEMBL DAT file:

http://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trEMBL.dat.gz

Description

This script imports PTM annotations into the PTMDB from UniProtKB files. It also imports protein sequences, accessions and sequence version numbers.

Program: org.drd20.bioinformatics.database.ptmdb.Phospho.ELM (class)

Input: Location of downloaded Phospho.ELM flat file

(latest version: <http://phospho.elm.eu.org/dataset.html>)

Description

This script imports PTM annotations from a dump of the Phospho.ELM database. Note that annotations are excluded if the provided sequence version number and/or sequence do not match that of the same UniProtKB entry stored in the PTMDB.

Program: org.drd20.bioinformatics.database.NotPtm (class)

Input: Location of downloaded Swiss-Prot DAT file

(latest version: see above)

Description

The original UniProtKB parser was not designed to extract negative PTM annotations. This script should be used to carry out this procedure until support is added to the main parser

Program: org.drd20.bioinformatics.database.ptmdb.GlycoSciences.importPDB2LINUCS.pl

Description

This script automates the process of importing glycosylation annotations into the PTMDB from the PDB using the PDB2LINUCS tool. This script uploads PDBs from a local PDB mirror to the PDB2LINUCS web server and then subsequently classifies the extracted glycans.

Program: org.drd20.bioinformatics.alignment.copao.CoPaO.pl

Input: The location of the two FASTA files that you would like to detect orthologues between.

Output: Outputs orthologous clusters, in the file format shown below, to the file results.csv

```

CLUSTER\tclusterId
SCORE\tscoreId
AORTH\tUniProtKB Accession (one or many)
BORTH\tUniProtKB Accession (one or many)
APARA\tUniProtKB Accession\tPrimary[0|1]\tConfidence value (zero or many)
BPARA\tUniProtKB Accession\tPrimary[0|1]\tConfidence value (zero or many)

```

Description

This script detects orthologues and in-paralogues between two lists of proteins in FASTA file format using the InParanoid algorithm. Note that a Java version of this program has also been created found in the same directory.

Program:

org.drd20.bioinformatics.database.ptmdb.HoBPret.Pfam.New.PfamAnnotationTransferMultiRun.pm

Input:

Pfam version [A|B]
 Restart flag [1|0]
 Run name
 Location of a file that lists TrEMBL proteins that should be included.

Output:

For each Pfam domain a directory is created in the output directory \$ENV{HOME}/GridBox/HoBPret/. This directory contains a number files that can be used to debug the cross-annotation procedure. Cross-annotation results are stored in the file results.csv in the format shown below.

Target	P35361	N-linked (GlcNAc...)	193	Query	Q6IEX5	N-linked (GlcNAc...)	148	131	14.29	BENCHMARK	NEW
	P35361-64-328				FGWSRYVPEG*N-----LTSC						
	Q6IEX5-19-265				LL-TFQLPFC*NAQVIDHYFCD						

Description

This program is responsible for carrying out the cross-annotation process. Cancelled cross-annotations runs can be resumed by setting the restart flag to true. This program will ignore Pfam domains that are marked as being complete in the output directory. Note that the output directory is statically set to \$ENV{HOME}/GridBox/HoBPret/. A new C++ version of this program is currently in development.